Classifiers as a model-free group comparison test

Bommae Kim¹ · Timo von Oertzen²

Published online: 3 April 2017
© Psychonomic Society, Inc. 2017

Abstract The conventional statistical methods to detect group differences assume correct model specification, including the origin of difference. Researchers should be able to identify a source of group differences and choose a corresponding method. In this paper, we propose a new approach of group comparison without model specification using classification algorithms in machine learning. In this approach, the classification accuracy is evaluated against a binomial distribution using Independent Validation. As an application example, we examined false-positive errors and statistical power of support vector machines to detect group differences in comparison to conventional statistical tests such as *t* test, Levene's test, K-S test, Fisher's *z*-transformation, and MANOVA. The SVMs detected group differences regardless of their origins (mean, variance, distribution shape, and covariance), and showed comparably consistent power across conditions. When a group difference originated from a single source, the statistical power of SVMs was lower than the most appropriate conventional test of the study condition; however, the power of SVMs increased when differences originated from multiple

sources. Moreover, SVMs showed substantially improved performance with more variables than with fewer variables. Most importantly, SVMs were applicable to any types of data without sophisticated model specification. This study demonstrates a new application of classification algorithms as an alternative or complement to the conventional group comparison test. With the proposed approach, researchers can test two-sample data even when they are not certain which statistical test to use or when data violates the statistical assumptions of conventional methods.

Keywords Group comparison · Classifiers · Support vector machine · K-fold cross validation · Independent validation

Introduction

It is frequent in research to compare two groups in order to learn about differences between them (e.g., Griffiths et al. 2004; Howlin et al. 2000; Kanagawa et al. 2001; Sabbagh et al. 2006; Yang et al. 2000). For example, comparisons could be made between samples of men and women, old and young, or those exposed to control or treatment conditions. Although group comparison can be done in various data features (such as central tendency, spread, distribution shape), let's consider a case of central tendency by comparing two group means of a single variable. A *t* test is the most frequently used test to compare two means and draw a statistical conclusion about the group difference in many disciplines. When an observed group difference is not likely to happen by chance under a null hypothesis that there is no group difference (i.e., the probability of obtaining the mean difference by chance is lower than a significance level α), we interpret this result that the two group means are not the same.

The views expressed in this paper are those of the author(s) and do not necessarily reflect the position of the Federal Reserve Bank of Kansas City or the Federal Reserve System.

Parts of this manuscript were developed while the first author was at the University of Virginia

✉ Bommae Kim
bommae.kim@kc.frb.org

¹ Federal Reserve Bank of Kansas City, Kansas City, USA

² University of the Federal Defense Forces, Neubiberg, Germany

Statistical methods rely on an assumption that a model of interest is correctly specified. The traditional t test, for example, has two assumptions: a normal distribution and equal variances (homoscedasticity). If our sample violates either of them, the results may no longer be valid. Violation of the statistical assumptions can cause an inflated alpha error (false-positive error, type I error), falsely rejecting a null hypothesis when there is no true effect; or a decreased statistical power, correctly rejecting a null hypothesis when there is a true effect (Curran et al., 1996; Flora & Curran, 2004; Hu et al., 1992). In a hypothesis test, statistical power is the probability to find an effect when it exists. A statistical test with a high power is desirable so that it can detect an effect if there is one. Power is mainly a function of effect size and sample size (Rossi, 2013) and higher power can be obtained by larger effect size, larger sample size, or both.

On the other hand, an alpha error occurs when a statistical test finds an effect when there is no true effect. It is crucial to have an alpha error rate as designed in hypothesis testing. If we find a statistically significant effect at a significance level alpha .05, we interpret that the effect is unlikely to happen by chance because the probability of the effect under the null distribution is smaller than 5%. If the statistical test does not maintain the alpha error rate correctly and produces a higher alpha error rate than designed (alpha inflation), one cannot draw a proper conclusion based on the test. For example, if a test produces 20% of alpha error rather than 5% as set by the predetermined significance level, researchers who investigate new clinical treatment might incorrectly conclude that the observed effect is statistically significant (that is, the probability that the observed effect happens by chance is less than 5%) even though this finding could in fact happen by chance, as often as one out of five times. Alpha inflation is often mentioned in multiple comparisons with suggestions of correction (e.g., Bonferroni correction; Dunn 1961), but alpha inflations in other cases have not been recognized enough. Researchers could draw a doubtful conclusion if they are not aware of alpha inflation in the test; or, if they are aware of alpha inflation in the test, they would not be able to draw a conclusion because they cannot be certain about the true effect.

Fortunately, many alternative methods are developed in case of violation of statistical assumptions. For example, a non-parametric test equivalent to the independent is the Mann–Whitney U test in which medians are used instead of means. Such robust methods show higher statistical power than standard methods when statistical assumptions, such as a normal distribution, are violated (Erceg-Hurn & Mirosevich, 2008).

Means are not the only interests of group comparison. One may be interested in comparing variances, rather than means, of a single variable between two groups. It may be

important to be sure that a new drug shows an effect in the same way among adults and children. In such a case, one can use Levene's test (Levene, 1960). If the goal is to compare a relationship of two variables (correlation) between groups, one can use Fisher's z -transformation (Fisher, 1915). There are many methods to test group differences and their robust versions are also developed (mainly for normality violation). Bootstrapping is also an option to empirically find the confidence intervals for test statistics.

However, all of these methods still assume a correct model specification on the origin of group difference. In other words, we should be able to correctly specify where a group difference may come from and use an appropriate method for it. If we do not have a priori information about where differences may exist, we do not have a way to examine if there is any difference between groups.

One may be interested in detecting any group difference but may not know (or may not be interested in) from where the differences come. With an advance of information technology, new types of data are now available. Examples include Internet user behavior (e.g., page view history and language usage in social media), consumer behavior (e.g., credit card transaction records), and mobile data (e.g., phone call logs and GPS information). These massive data contain unique information about people's behavior. It is not hard to imagine that it is difficult to specify a model using such new types of data. On the other hand, the goal may be to avoid alpha inflation (inflated false-positive errors) for any reason. In clinical trials, we may not know how a new chemical compound could affect a subject and it may be crucial not to draw a presumptive conclusion of positive effects based on uncertain models.

As discussed, many statistical methods are available for misspecification of distribution such as non-normality (Hollander et al., 2013; Wilcox, 2012), however misspecification of model (from where differences come) has not been studied enough. In this article, we will suggest a new approach of group comparison without model specification.

A new approach of group comparison using classifiers

For data with an unknown model structure, classification algorithms in machine learning are an alternative to model-based tests. Classification algorithms do not require model specification and are widely used in other areas where a model or a distribution is rarely defined, such as pattern recognition (Jain et al., 2000; Mohammed et al., 2011; Rosten et al., 2010), bioinformatics (Che et al., 2011; Inza et al., 2010; Saeys et al. 2012; Upstill-Goddard et al., 2013), and language processing (Indurkha & Damerau, 2012; Ganapathiraju et al., 2004; Sha & Saul, 2006). A main task of classification algorithm is to categorize data into groups based on their features. Clustering categorizes unlabeled

members into a small set of groups that show similar characteristics among group members. While clustering is an *unsupervised* approach when true group membership is not known, a classifier is an *supervised* approach when group memberships are at least partially known, and the known membership information can be used to decide unknown members (i.e., classifying new members into pre-existing groups).

A supervised classification assigns a group membership to unlabeled data based upon known features after it learns group memberships and corresponding patterns in data. A classifier predicts group membership better than chance *only when there is a distinctive (therefore detectable) difference between groups*. In other words, if there is no detectable difference between groups, a classifier cannot predict group membership correctly, and its prediction accuracy will be no better than guessing. If it guesses randomly, it demonstrates approximately 50% accuracy in the case of two groups. If it assigns group membership significantly more accurately than chance, this indicates that there are differences between groups and the classifier has successfully learned the differences in observed features to decide group membership. The probability of classification accuracy follows a binomial distribution for two classes.

Classifiers usually do not have assumptions about model specification including data distribution, and can be useful if research mainly aims to know whether or not any difference exists between groups when we do not have sufficient knowledge to specify a model.

Support vector machine

Among various classification algorithms (see Kotsiantis 2007), support vector machine (SVM; Cortes & Vapnik 1995; Vapnik 1998) is one of the most widely used machine learning methods because of its simplicity of use and flexibility with different tasks (Bennett & Campbell, 2000), and has shown excellent performance in many applications (Wang & Pardalos, 2015), such as computer vision (Drucker et al., 1999; Han & Davis, 2012; Mohammed et al., 2011; Osuna et al., 1997; Rosten et al., 2010), bioinformatics (Brown et al., 2000; Che et al., 2011; Inza et al., 2010; Saeys et al. 2012; Upstill-Goddard et al., 2013), and fMRI analysis (Poldrack et al., 2009; Serences et al., 2009; Wang et al., 2007), geosciences (Li et al., 2012; Mountrakis et al., 2011; Pradhan, 2013), and finance and business (Huang, 2012; Yang et al., 2011).

An SVM finds a linear decision surface (hyperplane) that can separate two classes with the largest distance (margin) between borderline observations (support vectors). If such a linear decision surface is not found in the original space (input space), the data is mapped into a higher dimensional space (feature space) where the separating decision surface

exists. The feature space is constructed by mathematical projection (kernel trick). The details of SVM would be beyond the scope of this paper, and there are many resources that cover the SVM in depth (Vapnik, 2000; Cristianini & Shawe-Taylor, 2000), and Noble (2006) provides a short introduction to SVM without using technical terms. In brief, SVM finds the best multi-dimensional separation based on any available variables through training and the separation is used to predict a group membership of an unknown member.

Although SVM is used in the current study, the underlying idea is the same for other classification methods. We could apply this framework using regularized logistic regression, linear discriminant analysis (LDA), regression trees (CART; Breiman et al. 1984), or neural networks (Garson, 1998).

Performance of classifier

If a classifier predicts group membership more accurately than chance, it is reasonable to conclude that there are group differences that the classifier is able to detect. The simplest way to assess the prediction accuracy of classification is to split data into two sets (a training set and a test set) and to test classifier performance with the test set after training a classifier with the training set (i.e., learning patterns in the training data). Although this is straightforward and simple, a classification performance is likely to suffer due to the smaller sample size of training after splitting data into two.

A more common method to assess classification accuracy is *k*-fold cross-validation (CV; Kohavi et al. 1995). It splits data into *k* partitions; a classifier tests its performance with the data in one of the partitions after trained with the data in *k*-1 partitions and it repeats the test and training process for every partition. CV shows higher power with larger *k* because it uses a larger training set, which allows the classifier to learn the data pattern better, while always keeping the total test set size the same. If the number of partitions *k* is the same as the total sample size *N* (i.e., *k* = *N*, testing with one element after training with everything else), it is called leave one out (LOO); this method is often maximizing the power of the classifier to detect an effect if any is present (Kohavi et al. 1995).

A two-class classification accuracy follows a binomial distribution. However, it has been found that CV shows inflated alpha error (Dietterich, 1998; Salzberg, 1997) due to dependency in repeated testing and training (von Oertzen & Kim, [under review](#)). An alternative validation method, Independent Validation (IV) has been suggested by von Oertzen and Kim ([under review](#)) to measure classifier performance. IV uses an element for training only after using it for a test to remove dependency in testing and training. Although it shows less power than LOO because it has a smaller training

size, the IV shows no alpha inflation. This study adopts IV to measure classification accuracy and to obtain test statistics.

In this article, we will show how a classifier can be used as a group comparison test and we will show its performance in comparison to widely used conventional tests. We will first examine univariate data and expand our study to multivariate data.

Univariate data

Method

We ran Monte Carlo simulations with a single variable to illustrate how SVMs can be applied in group comparison. Group differences were made in means, variances, distribution shapes (kurtosis), or combinations of them. Figure 1 shows how two sample distributions appear in the study conditions. We compared statistical power and alpha error rate of SVM and conventional tests. For the conventional tests to compare with, we used *t* test (homogeneity assumed), Levene's test (centered at means), and Kolmogorov–Smirnov test (K-S test; Massey 1951), which were most widely used for group comparison of means, variances, and distributions, respectively.

The K-S test is a non-parametric test, which compares two single-variable distributions. The K-S test can detect differences not only in distribution shapes (e.g., skewness

and kurtosis) but also in means and variances of two single-variable samples. We included this method to investigate the usefulness of classifiers as a universal assumption-free two-sample comparison test.

Each simulation concluded that a method detected an effect (i.e., a group difference), if a *p* value was smaller than a significance level alpha. The significance level was set at .05 in this study. Statistical power was computed as a percentage of significant results (i.e., $p \leq .05$) over total replications when a group difference truly existed (group difference > 0). An alpha error rate was computed as a percentage of significant results over total replications when there was no group difference in truth (group difference $= 0$).

For data generation, group 1 data were simulated as $x \sim N(0, 1)$. Group 2 data were simulated by adding a mean difference, adding a variance difference, combining two normal distributions with different variances, and combinations of these. The sample size for each group was 100, resulting in 200 in total. Each condition was replicated 1000 times. All computation was done in R (version 3.3.1; Core Team 2016) and an R package 'e1071' (version 1.6-7; Meyer et al. 2015) was used for SVMs.

For the best results of classification, it is a common practice to tune parameters in machine learning. In this simulation study, we did not tune parameters as thoroughly as we were supposed to do in order to save computation time. We used only a Gaussian kernel, and tuned a $\gamma \in$

Fig. 1 Illustration of two sample data. The *dashed line* represents a normal curve and the *solid line* represents study conditions. **a** Difference in means. **b** Difference in variances. **c** Difference in distribution shapes. **d** Differences in means, variances, and distribution shapes simultaneously

{0.001, 0.01, 0.1, 1} of a Gaussian kernel and a soft-margin cost $C \in \{1, 10, 100, 1000\}$. The best parameters were selected among all combinations of γ and C using tenfold cross-validation. In the real application, researchers should search best parameters more thoroughly to obtain the best classification results.

Statistically significant results of SVM were based on the accuracy of group membership prediction. If an SVM predicts group membership of data more accurately than by chance, it can be interpreted that an SVM detects group differences because two groups are somehow different from each other. As discussed earlier, the classification accuracy of SVM was measured by IV (von Oertzen & Kim, [under review](#)) to maximize its power while not introducing alpha inflation.

The IV procedure takes an initial training set of data, which is never used for testing, and a test size, with which a classifier is tested at each validation. Once a classifier is trained using the initial training set, the classifier is tested using a test set of the test size, randomly chosen from the rest of data. The used test set is combined to the training set for the next validation, and the classifier is trained using the increased training set and tested using a new test set. Therefore, each validation is completely independent in the IV procedure.

In the current study, an initial training size was 20 (ten from each group), and a test size for each validation was two (an instance from each group). An SVM was trained with the 20 training instances and tested with the two test instances in the first validation. The two test instances were then combined to the training set, increasing the training set by two each time. The validation process ended once the rest of data were tested. As a result, the SVM was tested

for the total number of instances subtracted by the initial training set size. For example, the total test set size is 180 for $N_{total} = 200$. A p value of the method was computed from the number of correct prediction under a binomial distribution.

Results

The study results for alpha error and statistical power of each method and the simulated data characteristics of group 2 are summarized in Table 1. When there was no difference between groups (condition 1), all methods kept alpha errors (type I error; false-positive rate) roughly at the desirable level 5%. As shown in Table 1, the alpha error of SVM was seemingly much lower than .05. This is because a binomial distribution is not continuous when a sample size is small. For $N = 200$, the smallest number for $p \leq .05$ is 112 at .038, which is the exact result in this study. The next smallest is 111 at .051, which exceeds the significance level .05; therefore the cut-off was set at 112. Alpha errors of SVM will be closer to .05 with a larger sample size.

The study condition to create group differences in means, variances, and distributional shapes were specifically chosen so that the corresponding conventional tests (t test, Levene's test, and K-S test) achieve approximately 30% of statistical power for M only, SD only, shape only conditions, respectively. As expected, the t test and Levene's test showed the highest powers when a difference existed only in means or variances, respectively (conditions 2, 3). Unexpectedly, the SVMs achieved a higher power than the K-S test when a difference existed only in distribution shapes (kurtosis), even though the K-S test is best known to compare two distribution shapes (condition 4).

Table 1 Statistical power of t test, Levene's test, K-S test, and SVM to detect group differences

Condition		Method				Data characteristics of group 2			
(Difference in)		t-test	Levene	K-S	SVM	Mean	SD	Kurtosis	Skewness
1	None	0.06	0.06	0.04	0.04	0.00	1.00	0.00	0.00
2	M	0.31	0.07	0.19	0.13	0.21	1.00	0.00	0.00
3	SD	0.05	0.29	0.05	0.09	0.00	1.17	0.00	0.00
4	Shape	0.05	0.21	0.32	0.60	0.00	1.00	2.42	0.00
5	M, SD	0.27	0.30	0.22	0.15	0.21	1.17	0.00	0.00
6	SD, Shape	0.05	0.02	0.16	0.45	0.00	1.17	2.42	0.00
7	M, Shape	0.31	0.21	0.52	0.67	0.21	1.00	2.36	0.29
8	All	0.28	0.03	0.38	0.55	0.21	1.17	2.37	0.25

Note 1. Group differences were made in a way to achieve approximately 30% power in M only, SD only, and shape only conditions for t test, Levene's test, and K-S test, respectively

Note 2. Group 1 $\sim N(0, 1)$, resulting in M = 0, SD = 1, kurtosis = 0, skewness = 0. $N_{total} = 200$ with an equal group size

Note 3. Condition 1 shows alpha errors at the significance level .05

Note 4. Group 2 data characteristics were computed from N = 1,000,000

The study condition to create group differences in means, variances, and distributional shapes were specifically chosen so that the corresponding conventional tests (t test, Levene's test, and K-S test) achieve approximately 30% of statistical power for M only, SD only, shape only conditions, respectively. As expected, the t test and Levene's test showed the highest powers when a difference existed only in means or variances, respectively (conditions 2 and 3). Unexpectedly, the SVMs achieved a higher power than the K-S test when a difference existed only in distribution shapes (kurtosis), even though the K-S test is best known to compare two distribution shapes (condition 4). The t test showed consistent behavior in all conditions. It was able to detect when group differences happened in means regardless of variances and distribution shapes.

On the other hand, the Levene's test produced peculiar results when differences occurred in distribution shapes. The Levene's test, which is known to compare two sample variances, detected differences in distribution shapes too, even though their variances were equal (conditions 4, 7). This implies that a Levene's test as a homogeneity test fails if distribution shapes are not equal between groups. Moreover, the Levene's test detected group differences well in SD only and shape only (conditions 3, 4), but it showed extremely low performance when differences are made in both (conditions 6, 8; in fact, lower than the significance level). Despite that, the Levene's test was the best method if a difference happens only in variances, given that the K-S test and SVMs were unsuccessful at detecting group differences in SD only (condition 3).

The K-S test was able to detect any difference in two sample distributions. Note that the statistical power of K-S test in SD only condition was .05, which happened to be at the significance level and therefore it seemed that the K-S test did not detect the group difference in variances (Condition 3). However, that was an issue of statistical power: in Table 1, the K-S test showed only .05 due to its low power

to detect a difference in SD. In fact, Table 2 (in the next part of the study) shows that the statistical power of K-S test was .13 with a larger difference in SD. Compared to the K-S test, the SVM performance ranged from 60 to 270% to the statistical power of the K-S test. Although the SVMs showed a lower power than the K-S test in M only and M & SD (conditions 1, 5), the SVMs outperformed the K-S test in all other five conditions. Overall, the SVMs detected differences in distribution shapes best, and it performed very well when group differences originated from multiple sources.

As mentioned, the values of each condition to generate data were chosen so that the conventional tests could achieve approximately 30% in power. However, the distribution shape difference may have overwhelmed the other sources in this study. Therefore, we studied more conditions in which the SVMs showed about 30% power in M only, SD only, and shape only conditions. This change resulted in larger group differences in mean and variance and smaller group differences in distribution shape in simulated data. These additional results are summarized in Table 2.

As expected, the t test performed well when differences occurred in means. As before, the Levene's test detected differences in variances well only when no difference was made in distribution shapes (conditions 10, 12). Again, the SVMs detected a difference in distribution shapes better than the K-S test did (condition 11). With a larger mean difference than in the previous set of conditions, the K-S test outperformed the SVMs when there were mean differences; the SVMs outperformed the K-S test in other conditions. The statistical power ranged from 5.9 to 67.0% for the K-S test and 14.7 to 55.3% for the SVMs. Except the SD & shape condition (condition 13), where all the methods suffered, the SVMs performed better when group differences came from multiple sources than from a single source.

When group differences occurred in variances and distribution shapes together, the t test, of course, did not perform well because it is not applicable for other than mean

Table 2 Statistical power of t test, Levene's test, K-S test, and SVM to detect group differences

Condition		Method				Data characteristics of group 2			
(Difference in)		t-test	Levene	K-S	SVM	Mean	SD	Kurtosis	Skewness
9	M	0.69	0.04	0.49	0.30	0.35	1.00	0.00	0.00
10	SD	0.06	0.77	0.13	0.30	0.00	1.35	0.00	0.00
11	Shape	0.04	0.12	0.15	0.30	0.00	1.00	2.00	0.00
12	M, SD	0.53	0.78	0.55	0.48	0.35	1.35	0.00	0.00
13	SD, Shape	0.05	0.29	0.06	0.15	0.00	1.35	2.00	0.00
14	M, Shape	0.71	0.11	0.67	0.55	0.35	1.00	1.90	0.38
15	All	0.54	0.27	0.41	0.33	0.35	1.35	1.95	0.29

Note. Group differences are made in a way to achieve approximately 30% power in M only, SD only, and shape only conditions for SVMs

differences. Aside from the t test, all the methods that were supposed to detect group differences did not do so very well (condition 13). It appeared that the increased variance and increased kurtosis caused a counter-effect, rather than causing a higher magnitude of difference. To confirm this, we ran an additional simulation, where variance of group 2 was smaller than group 1, while kurtosis of group 2 was greater than group 1 as before. The t test, Levene's test, K-S test, and SVMs showed statistical power of .05, 1.00, .69, .89, respectively. The data characteristics of group 2 were $M = 0.00$, $SD = 0.65$, kurtosis = 2.00, skewness = 0.00.

Multivariate data

Classifiers also can be conveniently applied in multivariate data without model specification. Using the similar framework as before, this section further investigated classifiers' performance in a bivariate case and four-variable case. This study aims to shed light on performance of classifiers as a group comparison test in high-dimensional data.

Method

We examined multivariate data of two features and four features. In the two-variable case, group 1 data were simulated as $x \sim N_k(\mu, \sigma)$, where $k = 2$, $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Group differences were made in covariance, means, and variances. The group differences in mean and/or variance were simultaneously made in both variables of group 2. For example, mean scores of both variables were higher in group 2 than in group 1 by 0.30 each. The magnitude of difference between groups was chosen so that the SVMs achieve approximately 20% power. For comparison, Fisher's z -transformation (Fisher, 1915) was used to test if a correlation of two variables in group 1 was different from group 2.

We also examined four-variable data. While the differences were made in four variables rather than two, the magnitude of difference between groups was exactly the same as in the two-variable case. For comparison, MANOVA (multivariate analysis of variance) was used to test if mean differences among four variables were statistically significant between two groups.

As in the previous univariate case, the sample size of each group was 100, resulting in $N_{total} = 200$. As before, the SVMs was tuned only roughly and the IV computed p values for classification results. Each condition was replicated 1000 times to compute statistical power to detect group difference. An additional R package 'MASS' (version 7.3-45; Venables & Ripley, 2002) was used to simulate data from a multivariate normal distribution.

Results

The results are summarized in Table 3. In the two-variable case, the Fisher's z -transformation showed high power to detect group difference in correlations when the group difference originated from which the test is designed to detect (conditions 17, 20). Not surprisingly, it did not detect any group difference when differences did not involve covariances (condition 17, 18, 20). The performance of Fisher's z -transformation considerably decreased when group differences also existed in variances of data (conditions 21, 23). Consistent to the previous univariate case, the SVMs were able to detect group differences from any sources, yet with relatively low power. Again, the SVMs showed higher statistical power when group differences came from multiple sources than from a single source.

In the four-variable case, the statistical power of SVMs to detect group differences was substantially increased from the ones of bivariate data in all conditions. The extent of improvement widely varied across conditions. Notably, it showed surprisingly strong performance when differences were made in covariances among variables (conditions 19, 21, 22).

As expected, the MANOVA performed better in M only (condition 17) and the SVMs did so in Var only (condition 18) and Cov only (condition 19). However, beyond our expectation, the SVMs showed higher statistical power than the MANOVA whenever differences came from more than a single origin, including the conditions where mean differences existed (conditions 20, 22, 23). Note that the Fisher's z , which is a test for correlation, outperformed the SVMs whenever there was a difference in covariance between two variables. MANOVA, a test for mean differences, did

Table 3 Statistical power of SVMs in multivariate data

Condition (Difference in)		2-variable		4-variable	
		Fisher z	SVM	MANOVA	SVM
16	None	0.04	0.03	0.06	0.04
17	M	0.05	0.21	0.73	0.36
18	Var	0.04	0.21	0.05	0.37
19	Cov	0.94	0.20	0.07	0.74
20	M, Var	0.06	0.41	0.61	0.65
21	Var, Cov	0.62	0.27	0.05	0.53
22	M, Cov	0.95	0.35	0.49	0.80
23	All	0.59	0.38	0.43	0.68

Note 1. Group differences are made simultaneously in all variables of group 2

Note 2. The magnitude of group difference of each condition is the same in two-variable and four-variable cases

so with four variables only when group differences were made in means but nothing else. In sum, the SVMs outperformed the MANOVA for all conditions except for M only (condition 17).

Discussion

In this study, a classifier was adopted as a group comparison test without model specification, producing a p value of how likely differences would exist between groups using Independent Validation (IV). SVMs were able to detect group differences regardless of their origins. The SVMs detected group differences even better than the K-S test did when a difference originated from distribution shapes. Among the methods examined in this study, the SVMs showed the most consistent performance across conditions. Moreover, SVMs' power improved when group differences came from multiple sources or when data contained multiple variables.

Although this study showed potentials of classifiers as a universal group comparison test, the study conditions were limited to simple cases. In real application, data are likely to consist of multiple variables and it is likely that group differences reside in multiple unknown sources including intertwined relationships from both observed and unobserved variables. Another limitation is that group sizes are rarely equal in real data and classification performance can be greatly affected when data are highly unbalanced.

An additional limitation important to note is that the proposed method is restricted to a single variable and it is not readily applicable to two or more independent variables and their interaction. However, classifiers can be applied to multiple classes (groups) and we could, in the future, extend this method to multiple variables. This is worthwhile research to pursue when moving forward.

As mentioned earlier, unbalanced data could affect classification results greatly. In addition, the simple classification accuracy would not be sufficient when group sizes are unequal; that is, we need to consider both precision and recall. In future research, it will be necessary to develop methods to evaluate results from unbalanced data.

In spite of these limitations, this study demonstrates a new application of classifiers. As an alternative or complement to the conventional statistical methods, classifiers can serve as a tool to detect any differences between groups. To that end, we have adopted IV in order to evaluate classification results and provide researchers with p values as in the conventional statistical tests.

Interestingly, the SVMs performed better finding differences in kurtosis than the K-S test, which is known as a tool to compare distribution shapes. By preprocessing data, classifiers may be further developed to serve as a test to

compare distribution shapes. Notably, the SVMs performed well when data departed from a normal distribution. Given that real data are hardly ever normally distributed in real data (Micceri, 1989), practical usefulness of classifiers is even more promising.

In this study, we clearly identified the origins of group differences. However, it is impossible to do so in a real-world setting. We never know from which group differences may come. It could be means, variances, and any higher order moments like kurtosis, or something we are not even aware of. Additionally, it is highly likely that group differences come from multiple sources simultaneously and the effects intertwine. It becomes practically impossible for researchers to specify all possible combinations. In such cases, classifiers can be a convenient tool to compare groups without model specification. As shown in the current study, if differences happen only in the way specified by researchers, conventional methods support their goal very well as designed. When that is not the case, classifiers can offer advantages over conventional tests.

Such model specification includes relationships between two groups. This study used two independent groups (between-subject design), but classifiers can apply to related groups too (within-subject design). The application of machine learning is very flexible because it does not make assumptions about the nature of data (e.g., independence between classes).

A linear relationship with a normal distribution is often assumed in data analysis. That assumption is especially prevalent in practice in the social and behavioral sciences. However, it might not be a reasonable assumption for new types of data that are brought by technological advancement. With machine-generated data (e.g., GPS locations and phone call logs), we may have difficulties in specifying models. For example, one may want to investigate relationships in GPS location data and compare those of a minority to a majority population. One may also want to investigate relationships of phone call length and text message frequency of both women and men, and compare those corresponding relationships to each other. With website visit history, researchers may want to investigate relationships of how often and how long users visit Internet websites and compare those from mobile phones to personal computers. Model specification, from a choice of probability to specification of relationships, would be a barrier to answer research questions using these non-traditional data.

SVMs, or most classifiers in machine learning, can be used as a model-free group comparison test. The extent of statistical modeling varies across statistical methods. Some methods require a higher extent of model specification, while some other methods less rely on researchers' model specification. For example, nonparametric approaches or resampling methods are greatly flexible in regard to some

statistical assumptions, such as data distribution. However, they still need researchers' model specification: where a group difference comes from.

In essence, traditional statistical methods intend to help researchers obtain information from data. Researchers specify a model based on a priori knowledge, therefore a model would not work as supposed if researchers' knowledge is insufficient or incorrect. In contrast, machine learning aims to extract information from data in hand without human intervention. Models are built based on given data and the modeling process relies minimally on researchers.

Although a classifier has the advantage of finding group differences from any sources without model specification, a disadvantage comes from that very advantage. Researchers do not know from where the group differences come when differences are detected. Further research is needed to establish a useful process once group differences are found. This includes narrowing down sources of differences by combining with different methods and establishing steps on how to interpret results and draw a meaningful conclusion.

We do not expect or intend to replace conventional methods with this new approach, however. We suggest that the proposed method can answer different questions, which more traditional methods may not. The classifiers can be used if a research question is to find any types of difference between groups; conventional methods are more useful if a research question is to find a specific difference between groups. Just like deciding a priority between false-positive errors and false-negative errors depending on questions to answer (e.g., for cancer diagnosis, false-positive errors are more acceptable than the other), a choice between this new method and the conventional methods depends on research questions.

It can be crucial to find any types of group differences particularly in clinical applications or program evaluations. For example, in medical experiments, researchers may want to have an additional assurance that control and treatment groups are the same in multiple aspects after randomization. Or, when random assignment is not possible (e.g., using pre-existing data or using demographic information as an independent variable), one may want to examine if two groups are likely the same except for an independent variable and dependent variables (i.e., to check results of matching).

Most importantly, we do not believe only one should be used over the other. They can be used together in a complementary way. For example, a combination of a traditional method and the proposed method can examine if an independent variable (manipulation) causes differences other than it is supposed to do. For instance, after a *t* test determines a mean difference between control and treatment groups, classifiers can examine if the manipulation have caused any unanticipated differences other than means, by

subtracting the means from each group. An investigation of effective combinations of the methods remains for future research tasks.

Classifiers have shown their usefulness with big data (Borders et al., 2005). With technological advances, social science has started to adopt new methods to collect data, such as social networks, GPS, ERP (Stahl et al. 2012), and fMRI (Lemm et al., 2011). Technology allows researchers to ask new research questions with newly available information. It also allows researchers to conveniently collect thousands of data points, which was hardly imagined few decades ago. The new data types will provide us with enormous amounts of information, but little is known about the nature of the data (such as distributions or models), and it might be useless or impossible to determine given our rapidly changing information technology. Classification can be a powerful and convenient statistical test for these new data types. In the current study, we investigated SVMs as one example of a classification method. There are also many other model-free classification methods, such as *k*-nearest neighbor, or decision trees (see Kotsiantis, 2007). Classification methods can work as a universal, robust test that is an alternative or complement to traditional statistical methods.

It should not be the standard practice in science to rely on doubtful assumptions without examination or to simply ignore the assumptions underlying statistical methods. On the other hand, it is not desirable to limit research questions due to lack of current knowledge about the nature of data. Alternative statistical tests will allow us to explore new research questions with newly available data types.

References

- Bennett, K. P., & Campbell, C. (2000). Support vector machines: Hype or hallelujah? *ACM SIGKDD Explorations Newsletter*, 2, 1–13.
- Borders, A., Ertekin, S., Weston, J., & Bottou, L. (2005). Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research*, 6, 1579–1619.
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Boca Raton, FL: CRC Press.
- Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97, 262–267.
- Che, D., Liu, Q., Rasheed, K., & Tao, X. (2011). Decision tree and ensemble learning algorithms with their applications in bioinformatics. In H.R. Arabnia, & Q.-N. Tran (Eds.) *Software tools and algorithms for biological systems* (pp. 191–199). New York, NY: Springer.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge, UK: Cambridge University Press.

- Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, 1, 16–29.
- Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7), 1895–1923.
- Drucker, H., Wu, D., & Vapnik, V. N. (1999). Support vector machines for spam categorization. *Neural Networks*, 10, 1048–1054.
- Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293), 52–64.
- Erceg-Hurn, D. M., & Miroseovich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, 63, 591–601.
- Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4), 507–521.
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9, 466–491.
- Ganapathiraju, A., Hamaker, J. E., & Picone, J. (2004). Applications of support vector machines to speech recognition. *IEEE Transactions on Signal Processing*, 52, 2348–2355.
- Garson, G. D. (1998). *Neural networks: An introductory guide for social scientists*. London, UK: Sage.
- Griffiths, M. D., Davies, M. N., & Chappell, D. (2004). Online computer gaming: A comparison of adolescent and adult gamers. *Journal of Adolescence*, 27, 87–96.
- Han, B., & Davis, L. S. (2012). Density-based multifeature background subtraction with support vector machine. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5), 1017–1023.
- Hollander, M., Wolfe, D. A., & Chicken, E. (2013). *Nonparametric statistical methods*. Hoboken, NJ: John Wiley & Sons.
- Howlin, P., Mawhood, L., & Rutter, M. (2000). Autism and developmental receptive language disorder—a follow-up comparison in early adult life. ii: Social, behavioural, and psychiatric outcomes. *Journal of Child Psychology and Psychiatry*, 41, 561–578.
- Hu, L.-T., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, 112, 351–362.
- Huang, C.-F. (2012). A hybrid stock selection model using genetic algorithms and support vector regression. *Applied Soft Computing*, 12(2), 807–818.
- Indurkha, N., & Damerau, F. J. (2012). *Handbook of natural language processing* Vol. 2. CRC Press: Boca Raton, FL.
- Inza, I., Calvo, B., Armananzas, R., Bengoetxea, E., Larrañaga, P., & Lozano, J. A. (2010). Machine learning: An indispensable tool in bioinformatics. In R. Matthiesen (Ed.) *Bioinformatics methods in clinical research, volume 593 of Methods in Molecular Biology* (pp. 25–48). New York, NY: Humana Press.
- Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 4–37.
- Kanagawa, C., Cross, S. E., & Markus, H. R. (2001). Who am I? The cultural psychology of the conceptual self. *Personality and Social Psychology Bulletin*, 27, 90–103.
- Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI* (Vol. 14, pp. 1137–1145).
- Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 31, 249–268.
- Lemm, S., Blankertz, B., Dickhaus, T., & Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *NeuroImage*, 56, 387–399.
- Levene, H. (1960). Robust tests for equality of variances1. *Contributions to probability and statistics: Essays in honor of Harold Hotelling*, 2, 278–292.
- Li, C.-H., Kuo, B.-C., Lin, C.-T., & Huang, C.-S. (2012). A spatial-contextual support vector machine for remotely sensed image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 50(3), 784–799.
- Massey, F. J. (1951). The Kolmogorov–Smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253), 68–78.
- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2015). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-7.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156–166.
- Mohammed, A. A., Minhas, R., Jonathan Wu, Q., & Sid-Ahmed, M. A. (2011). Human face recognition based on multidimensional PCA and extreme learning machine. *Pattern Recognition*, 44, 2588–2597.
- Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3), 247–259.
- Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology*, 24, 1565–1567.
- Osuna, E., Freund, R., & Girosi, F. (1997). Training support vector machines: an application to face detection. In *Proceedings 1997 IEEE computer society conference on computer vision and pattern recognition, 1997* (pp. 130–136). IEEE.
- Poldrack, R. A., Halchenko, Y. O., & Hanson, S. J. (2009). Decoding the large-scale structure of brain function by classifying mental states across individuals. *Psychological Science*, 20, 1364–1372.
- Pradhan, B. (2013). A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Computers & Geosciences*, 51, 350–365.
- Core Team, R. (2016). R: A language and environment for statistical computing. Vienna, Austria.
- Rossi, J. (2013). Statistical power analysis. In I. B. Weiner, J. A. Schinka, & W. F. Velicer (Eds.) *Handbook of psychology: Research methods in psychology*, 2edn (pp. 71–108). Hoboken, NJ: Wiley.
- Rosten, E., Porter, R., & Drummond, T. (2010). Faster and better: A machine learning approach to corner detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 105–119.
- Sabbagh, M. A., Xu, F., Carlson, S. M., Moses, L. J., & Lee, K. (2006). The development of executive functioning and theory of mind a comparison of Chinese and us preschoolers. *Psychological Science*, 17, 74–81.
- Saeyns, Y., Wehenkel, L., Geurts, P., et al. (2012). Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. *Bioinformatics*, 28, 1766–1774.
- Salzberg, S.L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data mining and Knowledge Discovery*, 1, 317–328.
- Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychological Science*, 20, 207–214.
- Sha, F., & Saul, L. K. (2006). Large margin hidden Markov models for automatic speech recognition. In *Advances in neural information processing systems* (pp. 1249–1256).
- Stahl, D., Pickles, A., Elsabbagh, M., Johnson, M. H., Team, B., et al. (2012). Novel machine learning methods for ERP analysis: A validation from research on infants at risk for autism. *Developmental Neuropsychology*, 37, 274–298.
- Upstill-Goddard, R., Eccles, D., Fliege, J., & Collins, A. (2013). Machine learning approaches for the discovery of gene–gene interactions in disease data. *Briefings in Bioinformatics*, 14, 251–260.

- Vapnik, V. N. (1998). *Statistical learning theory* Vol. 1. New York: Wiley.
- Vapnik, V. N. (2000). *The nature of statistical learning theory*. New York, NY: Springer.
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*, 4th edn. New York: Springer. ISBN 0-387-95457-0.
- von Oertzen, T., & Kim, B. (under review). Independent validation remedies alpha inflation in classifier accuracy testing.
- Wang, J., Korczykowski, M., Rao, H., Fan, Y., Pluta, J., Gur, R. C., McEwen, B. S., & Detre, J. A. (2007). Gender difference in neural response to psychological stress. *Social Cognitive and Affective Neuroscience*, 2, 227–239.
- Wang, X., & Pardalos, P.M. (2015). A survey of support vector machines with uncertainties. *Annals of Data Science*, 1(3–4), 293–309.
- Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. Academic Press.
- Yang, N., Chen, C. C., Choi, J., & Zou, Y. (2000). Sources of work–family conflict: A Sino–US comparison of the effects of work and family demands. *Academy of Management Journal*, 43, 113–123.
- Yang, X.-S., Deb, S., & Fong, S. (2011). Accelerated particle swarm optimization and support vector machine for business optimization and applications. In *Networked digital technologies* (pp. 53–66) Springer.