How to improve the statistical power of the 10-fold cross validation scheme in Recommender Systems

Andrej Košir[*]
University of Ljubljana, Faculty
of Electrical Engineering
Tržaška cesta 25, Ljubljana,
Slovenia
andrej.kosir@fe.uni-lj.si

Ante Odić
University of Ljubljana, Faculty
of Electrical Engineering
Tržaška cesta 25, Ljubljana,
Slovenia
ante.odic@ldos.fe.uni-lj.si

Marko Tkalčič
Johannes Kepler University
Dep. for Comp. Perception
Altenberger Str. 69,
Linz,Austria
marko.tkalcic@jku.at

ABSTRACT

At this stage development of recommender systems (RS), an evaluation of competing approaches (methods) yielding similar performances in terms of experiment reproduction is of crucial importance in order to direct the further development toward the most promising direction. These comparisons are usually based on the 10-fold cross validation scheme. Since the compared performances are often similar to each other, the application of statistical significance testing is inevitable in order to not to get misled by randomly caused differences of achieved performances. For the same reason, to reproduce experiments on a different set of experimental data, the most powerful significance testing should be applied. In this work we provide guidelines on how to achieve the highest power in the comparison of RS and we demonstrate them on a comparison of RS performances when different variables are contextualized.

Keywords

recommender systems, evaluation, folding, paired testing, experimental design

1. INTRODUCTION

The discussions on the evaluation in recommender systems have been mostly focused on the choice of the evaluation metric (e.g. RMSE vs. precision etc.). However, very little focus has been given to the next step in evaluation: the comparison of two or more recommender systems (e.g. underlying algorithms or methods) using appropriate statistical tests (based on the chosen evaluation metric). In this paper we address the latter.

When comparing several recommender systems (for instance, method A_1 and method A_2), the task of determining which one is better is not trivial. There are two choices that the evaluators have to make: (i) choosing an appropriate evaluation metric m and, after that, (ii) choosing an appropriate comparison procedure. This is conceptually explained in Fig. 1. The vertical axis represents different scalar performance measures m (e.g. precision (P), precision at five (P@5) etc). Which one suits best the evaluation procedure depends on the domain specifics the evaluated RS is designed for. In Fig. 1 we assumed this is F measure indicated by the blue dot. Next, we select an appropriate significance test to answer the question whether the two methods are equivalent, i.e. $m(A_1) = m(A_2)$. Such test is the one for which all assumptions regarding tested data are meet and is the most powerful among such tests. In Fig. 1 we select the Wilcoxon signed rank test (indicated by the blue dot) since the resulting F measures typically do not meet the normality assumption of the t-test (which would be the most powerful choice).

Why do we need to apply significance testing here? Clearly, the right selection of a scalar measure is of crucial importance. However, when a stage of development of certain field reaches maturity (as is the case with RS today), the competing algorithms are getting closer and closer regarding their performances. Therefore, one can not tell whether the achieved performance gap $m(A_2) - m(A_1)$ is a result of true performance improvement or just a coincidence related to the data sample. To resolve the issue, a statistical significance test is applied to test the null hypothesis $H_0 : m(A_1) = m(A_2)$ against the alternative one $H_1 : m(A_2)! = m(A_1)$.

The statistical power of a significance test is the probability of rejecting a null hypothesis H_0 when it is not true (i.e. detecting the deviation from the null hypothesis). In most cases, the hypothesis claims that there is no effect, i.e. no association among variables or no difference among population means such as the performance of a RS measured by F-measure etc. This effect is thus the deviation from the null hypothesis. When the null hypothesis is not true, it means there is a real effect in the population. Therefore, the statistical power tells us how likely the test detects this real effect by rejecting the null hypothesis. For this reason, the statistical power is also called sensitivity in terms of how likely the test detects the effect, that is the deviation from the null hypothesis.

When the performances of the two methods A_1 and A_2 are close to each other but different, we wish to detect the difference in order to continue the development into the more promising direction. The power of significance testing is thus of crucial importance. Since statistical power, beside

[*]Corresponding author.

RepSys '13, October 12, 2013, Hong Kong, China.
Copyright 2013 ACM 978-1-4503-2465-6/13/10 ...$15.00.
http://dx.doi.org/10.1145/2532508.2532510.

the effect size, also depends on the sample size. And, since the user data acquisition required in the comparison of RS is difficult, we need to achieve the highest available statistical power for a given user test data. The problem how to achieve this is addressed in this paper.

High statistical power in RS comparison is also of crucial importance for the reproduction of experiments. The reproduction of a given experiment is executed on a different data set than the original one with a possibly different effect size. If one wishes to reproduce the result, the applied test must be powerful (sensitive) enough to detect the effect.

The issue of choosing the appropriate evaluation metric is an old one. Herlocker et al. [4] proposed a set of metrics to choose from, depending on the task at hand. A more recent comprehensive overview of metrics is given also in Shani [12]. There have also been attempts to capture multiple metrics, like business-wise, technical-wise and user-wise [11].

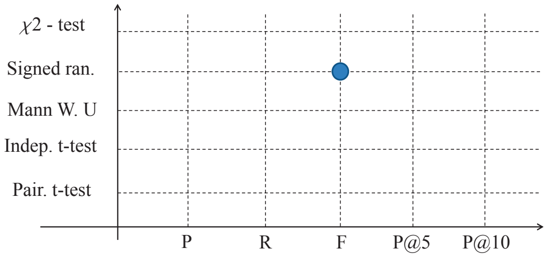

Figure 1: A conceptual representation of the two methods comparison selection. First we select a scalar performance measure such as F measure (horizontal axis) and then an appropriate significance test (vertical axis).

After the evaluation metric has been chosen it is applied to each of the observed recommender systems, which yields a set of measurements for each recommender system. To the best of the authors' knowledge, little discussion has been made with regards to the comparison of recommender systems' performance. Jannach [5] suggests using ANOVA to perform pairwise comparisons. One could also resort to methods used in machine learning (e.g. Demšar [3]). However, the choice of the appropriate comparison test depends heavily on the metrics chosen in step (i).

In this paper we address the issue of performing a correct comparison between two recommender system when the evaluation metrics is the confusion table of the classifier. More specifically, we address the pair-wise equivalence of the 10-fold cross validation scheme.

1.1 Problem statement and proposed solution

The problem we address is the lack of well-established guidelines for selecting the best statistical test when comparing two recommender systems. We propose guidelines for selecting the statistical test with the highest statistical power in a common scenario: the outcome of each of the recommender systems under observation is the set of confusion matrices yielded by the 10-fold cross validation scheme. The matter is particularly relevant in RS comparison since users' data is typically sparse and consequently, the statistical power of significance test is lower.

Furthermore, we present the outcomes of the proposed methodology on a set of sub-variants of a recommender system (based on the matrix factorization algorithm) with different contextual variables.

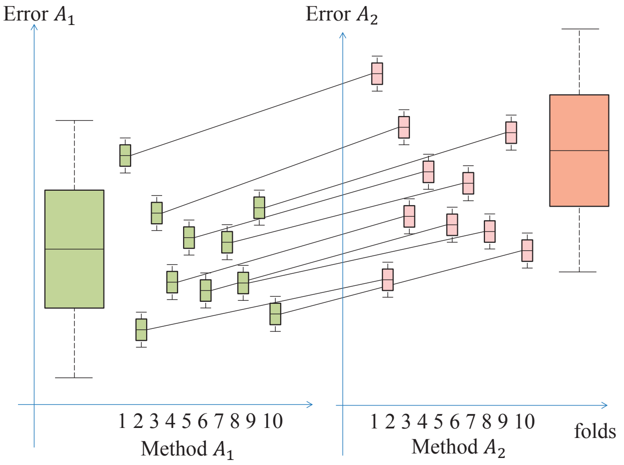

Figure 2: Illustration of paired vs. independent comparisons. The vertical axis represents the generic error of the analyzed method. The green (left) boxes represent the method A 1 and the red ones (right) represent a method A 2 . Each of the small boxes represents the one fold error values. A pair of small green and red boxes connected by a line represents the comparison of both methods in a selected fold indicated on the horizontal axis. Observe how the pairwise comparisons are significantly different while the aggregated is not.

2. SELECTING THE STATISTICAL TEST

In this subsection we list and briefly discuss the options available in the comparison of two methods in RS based on the 10-fold cross validation scheme and argue that the so called paired version in the one to select.

2.1 Criteria for significance test selection

In general statistical practice, a significance test is selected according to the following two competitive directions. They both have the same goal which is getting the most statistical power in the testing procedure. Typically, there are several statistical tests available that test roughly the same null hypothesis and they are ordered according to their power [8]. The first guideline that one must follow is to meet the assumptions of the significance test regarding the variable type (nominal, ordinal, numeric) and the tested data distributions. For instance, having numerical data, a paired t-test could be appropriate but if the data is not distributed normally, we have to select the Wilcoxon signed rank test. Any test for which the assumptions are met could be selected. The second guideline to follow is to select the most powerful test among acceptable ones. For instance, if the Wilcoxon signed rank test could be applied, the χ^2-test could also be applied but that would not be an optimal choice since it yields less statistical power than the Wilcoxon signed rank test.

The selection of the significance test also depends on the pre-processing of tested data. In our case this means the way we store and manipulate results of each fold comparison among methods A_1 and A_2. The guideline is as follows: if it is possible to apply a paired significance testing, one should select it (since it yields higher statistical power compared to the independent comparison). Usually the paired vs. independent choice is decided by the nature of the experiment and there is no way to perform a paired test if the data is independent (e.g. the compared values are measures on different subjects). On the other hand, if the data is paired one can perform an independent test but some statistical power is lost. Therefore, such selection would be a flaw in the experimental design. Such a design flaw can occur in the case of significance testing based of the 10-fold cross validation scheme as indicated in Fig. 2.

2.2 Paired vs. Independent Signific. testing

In this section, we list and discuss several options on how the 10-fold cross validation scheme-based comparison of two competitive methods, based on scalar performance measures and significance testing can be implemented. We also comment on each of them regarding the achieved statistical power.

We are comparing the performance of two competing methods A_1 and A_2 measured by a scalar performance measure $m(A_1)$ and $m(A_2)$. We test the null hypotheses $H_0 : m(A_1) = m(A_2)$ against the alternative one $H_1 : m(A_1)! = m(A_2)$. There are the following paths of reasoning:

1. **Same vs. separate folding**: one can (i) run the 10-folding scheme once and compare the competing methods on the same training / testing folds or (ii) run the folding scheme twice, first to estimate $m(A_1)$ and then to estimate $m(A_2)$ (with a different set of folds). The first option (same folding) typically yields higher power in testing the null hypothesis H_0 than the second one does. The reason is in the fact that in the case of independent testing the differences among compared values may cancel out while they are preserved when paired comparisons are made;

2. **Paired vs. independent testing:** in the case that we used the same-folding approach (first option of the previous step) there are two options on how to compare the two competing methods A_1 and A_2. First, one can perform paired testing where performance measures $m(.)$ of the two methods are compared on the same fold in a paired way (we call it the ProcPaired approach) and second, all performance measures $m(.)$ of each methods are grouped together and then compared as independent sets (we call it the ProcIndep approach). The situation is depicted in Fig. 2 and further discussed below. As known from the theory of hypothesis testing [8], paired tests are stronger than independent ones.

We claim that the ProcPaired procedure (using same folds in comparison and comparing rating predictions on the same test fold items) yields higher statistical power than the other options listed above. An outline of the recommended procedure is given at the end of this section.

It might seem obvious that one should always use ProcPaired when comparing two methods in RS. However, typically one confusion matrix is computed for each evaluation fold for each of the competing methods, these matrices are then summed to one confusion matrix for each method and, finally, a scalar measure such as F measure is computed and compared. Such a procedure is actually the ProcIndep procedure that yields to a lower power than available.

2.3 The recommended testing procedure

In this subsection we outline the testing guidelines ProcPaired that we recommend. It is based on the above given reasoning, also refer to Fig. 1. Recall that the scalar performance measure of the method A (such as F-measure) was denoted by $m(A)$. The procedure is as follows:

1. Select the scalar comparison measure (such as precision or F-measure);

2. Store the results of each fold and each method separately;

3. According to the specific features of the evaluation results (distributions etc.) select the most powerful test that meets these specific features of evaluation results (i.e. t-test for normally distributed numerical values etc.);

4. Perform the **paired** version of the selected test.

3. MATERIALS AND METHODS

In this section, we present the experimental design, data and experimental results in order to demonstrate the performance of the proposed procedure in Sec. 2.

3.1 Dataset

For the purposes of this work we have used the *Context Movie Dataset* (LDOS-CoMoDa), that we have acquired in our previous work [10].

Table 1: Contextual factors in the LDOS-CoMoDa dataset.

Contextual variable	Description
time	morning, afternoon, evening, night
daytype	working day, weekend, holiday
season	spring, summer, autumn, winter
location	home, public place, friend's house
weather	sunny/clear, rainy, stormy, snowy, cloudy
social	alone, partner, friends, colleagues, parents, public, family
endEmo	sad, happy, scared, surprised, angry, disgusted, neutral
dominantEmo	sad, happy, scared, surprised, angry, disgusted, neutral
mood	positive, neutral, negative
physical	healthy, ill
decision	user's choice, given by other
interaction	first, n-th

We have created an on-line application for rating movies which users are using in order to track the movies they watched and obtain recommendations (*www.ldos.si/ recommender.html*). Users are instructed to log into the system after watching a movie, enter a rating for a movie and fill in a simple questionnaire created to explicitly acquire the contextual information describing the situation during the consumption.

The part of the dataset used in this study consists of 1611 ratings from 89 users to 946 items with associated contextual factors. Additional information about our *Context Movies Database* (LDOS-CoMoDa) can be found in [7] and [10].

All the contextual factors and conditions acquired are listed in Tab. 1.

3.2 Experimental Design

The experimental design applied in this study is equivalent to the one used in our previous work [10]. It considers the improvement of a matrix factorization (MF) [6] based RS when different potentially contextual variables are contextualized in a MF model.

We used the 10-fold cross validation scheme on the data presented in Subsection 3.1. We compared the procedures ProcPaired and ProcIndep detailed in Subsection 2, the results are given in Table 2.

4. RESULTS

The experimental results on the comparison of the three pairs of methods are reported in this Section. In Table 2 we report the p-values and the achieved (post hoc) statistical powers. The statistical power was computed as suggested by Cohen [2] using the free tool [1]. The listed p values were computed using the Wilcoxon signed rank test [13] for paired comparisons and the Mann Whitney U [9] test for independent ones.

Table 2: The achieved (post hoc) statistical power for the paired test (pw pa.) and for the independent test (pw in.) along with the computed p-values using the Wilcoxon signed rank test and the Mann Whitney U test, respectively.

Id	Var1	Var2	pw pa.	p pa.	pw in.	p in.
1	physical	weather	0.42	0.001	0.14	0.24
2	decision	social	0.997	0.004	0.25	0.19
3	interaction	social	0.06	<0.01	0.05	0.43

Observe that the achieved post hoc statistical power is higher in the proposed (paired) procedure as in the case of independent ones.

4.1 Discussion

The first combination (physical vs. weather) exhibits moderate to low power in the paired version (pw=0.42) and a very low power in the independent version (pw=0.14). An increase in the test power is relevant since the power of 0.14 is not useful while the power 0.42 is low but still useful. The paired comparison detected the difference in the proposed methods' performances ($p < \alpha = 0.05$) while the unpaired version did not. This is a consequence of the low power of the unpaired test.

The second combination (decision vs. social) achieves a power close to 1 in the paired version and a very low power of 0.25 in the independent version. The difference in power is again substantial. Regarding the acceptance of the null hypotheses we have the same situation and the same explanation as in the case of the combination 1 (previous paragraph).

The third combination (interaction vs. social) achieves an extremely low power in both versions, far below the values that allow an interpretation (this is due to the very low effect size, we do not discuss it here). However, the paired version rejects the null hypothesis ($p < 0.01$, indicating that the methods are of different performances) despite the extremely low power while the independent version does not reject it. Note that despite the low achieved power of the paired version the rejection of the null hypotheses is valid.

5. CONCLUSION AND FUTURE WORK

In this study we outlined the procedure for the comparison of two methods in RS based on the 10-fold cross validation scheme that achieves high statistical power. We demonstrated the statistical power improvement on real users' data in comparison of matrix factorization models with different contextualized variables.

Further work will concentrate on a comparison of RS regarding the selected final tasks such as best five and not limited to scalar performance measures such as precision at five.

6. ACKNOWLEDGMENTS

This work was funded by the Slovenian Research Agency (ARRS) under grant number R-819 and was supported by the EU Seventh Framework Programme FP7 / 2007-2013 through the project PHENICX (grant no. 601166).

7. REFERENCES

[1] F. F. A. Buchner, E. Erdfelder and A. Lang. G*power (version 3.1.3). http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/, 2010.

[2] J. Cohen. Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum, 1988.

[3] J. Demšar. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res., 7:1–30, Dec. 2006.

[4] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 22(1):5–53, Jan. 2004.

[5] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. Recommender Systems: An Introduction. Cambridge University Press, 1 edition, 9 2010.

[6] Y. Koren, R. Bell, and C. Volinsky. Matrix Factorization Techniques for Recommender Systems. Computer, 42(8):30–37, Aug. 2009.

[7] A. Košir, A. Odić, M. Kunaver, M. Tkalčič, and J. Tasič. Database for contextual personalization. Electr. review, 78(5):270–274, 2011.

[8] E. L. Lehmann and J. P. Romano. Testing Statistical Hypotheses. Springer, 3rd edition, 11 2010.

[9] H. B. Mann and W. D. R. On a test of whether one of two random variables is stochastically larger than the other. Ann. of Math. Statistics, 18(1):50–60, 1947.

[10] A. Odić, M. Tkalčič, J. F. Tasič, and A. Košir. Predicting and detecting the relevant contextual information in a movie-recommender system. Interacting with Computers, 25(1):74–90, 2013.

[11] A. Said, B. J. Jain, and S. Albayrak. A 3d approach to recommender system evaluation. In Proceedings of the 2013 conference on Computer supported cooperative work companion, pages 263–266. ACM, 2013.

[12] G. Shani and A. Gunawardana. Evaluating recommendation systems. pages 257–298, 2011.

[13] F. Wilcoxon. Individual comparisons by ranking methods. Biometrics Bulletin, 1(6):80–83, 1945.