Chance influence in datasets with a large number of features

Abdel Aziz Taha, Alexandros Bampoulidis, Mihai Lupu
Research Studios Austria
Data Science
1090 Vienna, Austria

***Abstract*—Machine learning research, e.g. genomics research, is often based on sparse datasets that have very large numbers of features, but small samples sizes. Such configuration promotes the influence of chance on the learning process as well as on the evaluation. Prior research underlined the problem of generalization of models obtained based on such data. In this paper, we deeply investigate the influence of chance on classification and regression. We empirically show how considerable the influence of chance such datasets is. This brings the conclusions drawn based on them into question. We relate the observations of chance correlation to the problem of method generalization. Finally, we provide a discussion of chance correlation and guidelines that mitigate the influence of chance.**

***Index Terms*—Chance correlation, Generalization, Reproducibiliy, sparse data, Genomics**

I. Introduction

Datasets with very large numbers of features, but at the same time small numbers of samples are being frequently used in machine learning, especially in the medical domain and genomics research. A significant problem with this kind of data is the low reproducibility and generalizability of the results concluded based on it. For example, features reported by one research group as predictive regarding some outcome either widely differ from other groups' features or are not predictive when applied on other groups' data. [1] [2] [3].

Michiels et al. [4] reanalyzed seven studies that were claimed to be able to predict cancer using microarray data. They reported that the results of most of these studies are overoptimistic and five of them provide prediction methods that are not better than a random predictor. In particular, they reported instability in the feature selection in the sense that the features selected by the algorithms as predictive regarding the underlying outcomes significantly change depending on the patients considered in the training sets, such that the feature selection can be described as unrepeatable. Gene lists found by different research groups typically have only a small overlap. They also reported that this instability decreases with increasing number of samples used for training and evaluation. Jianping et al. [5] emphasized the great influence of the ratio between the number of features and the sample size on the reliability and reproducibility of prediction.

This work was carried out under support of projects VISIOMICS (FFG COIN), Safe-DEED (H2020 825225) and DMA (FFG, BMVIT 855404).

The Motivation for this research was a remarkable observation while performing experiments on an RNA microarray dataset of 16000 features (genes) of 80 diseased neuroblastoma children. The task was to predict the survival time (time until an event, e.g. relapse or death). A prediction correlation of more than 97% was achieved using a simple regression model in a cross validation experiment after performing a simple feature selection. This high accuracy was doubtful, given the small number of samples. In a follow-up experiment, we replaced all the gene data with random numbers and kept the target (survival time) unchanged. We applied exactly the same feature selection and regression algorithms. The results of the trained prediction model obtained a correlation above 95%. This remarkable observation motivated a deeper look in the influence of chance on ML models. This observation as well as the literature denoting the problem of generalizability with this kind of datasets demonstrates the need for better understanding of chance influence.

In this paper, we first demonstrate the very large influence of chance correlation on training and evaluating prediction algorithms despite using the common cross validation. We show that these results can be confirmed using thousands of random datasets with different dimensionalities and different data types for both classification and regression. We also show that the way how feature selection is performed has a great impact on chance correlation. We provide discussion of the relation between chance and the dataset dimensionality including the number of classes. Finally we conclude by providing guidelines to mitigate the influence of chance on prediction models.

II. Related Work and background

Prior research aims at finding optimal configurations in terms of sample size as well as feature number. In this section we summarize some of this research.

Learning curves: To predict how the classification accuracy would change when the sample size is increased, it is common to use the learning curve modelling. A learning curve is a model that describes the progress of a learning process, e.g. the accuracy of a ML algorithm as a function of the number of examples fitted in the learning. A common method to implement a learning curve is to fit the inverse power law using small samples, [6], i.e.: $f(n) = an^{-\alpha} + b$, where f is

the learning rate for n samples and a, b, and α are parameters that depend on the algorithm and the dataset.

Many approaches follow this principle to predict the accuracy of an algorithm in a confidence interval around the learning curve, given a number of samples (Figueroa et al. [7]) or to estimate the minimum number of samples required for a classification to keep the error in a particular confidence interval (Mukherjee et al. [8], Dobbin et al. [9]). However, these approaches aim at optimizing the accuracy by finding the optimal number of samples and they do not consider the generalizability and the influence of chance correlation.

Bias regarding feature selection: Ambroise et al. [10] thoroughly discussed the bias caused by the feature selection method used prior to cross validation, when feature selection is performed on the entire dataset. They stated that in this case the estimation of prediction error is too optimistic, because the kind of testing is influenced by the bias stemming from the fact that the test set is a subset from set (in this case the entire set) used for feature selection. As bias correction, they suggested using a special cross-validation and bootstrapping method to correct the biased results.

Ein-Dor et al. [1] investigate the problem of robustness in the feature selection in genomics research, i.e. that genes identified as predictive regarding an outcome vary depending on the samples included in the training such that there is only a small overlap between gene lists identified using different training subsets. Their results show that thousands of samples are required to achieve an overlap of more 50% between the gene lists.

Chance Correlation: Kuligowski et al. [11] investigated the prediction accuracy in metabolomics using Partial Least Squares Discriminant Analysis. They reported that cross-validation after feature selection provides overoptimistic results due to the chance correlation. The effect of chance correlation is expressed by means of p-values calculated by using a permutation test including the variable selection, but they don't relate the chance to the number of features.

III. Notation

We provide definitions that hold for the whole paper. Two types of experiments, which will be used for analysis in this paper are defined in the following.

Definition 1: Regression based on random: Let $D = \{C_1, ..., C_m\} \cup \{C^*\}$ be a random dataset of the shape $m \times n$ where C_1 to C_m are columns (features) in the dataset and C^* is the target column. All values are of numeric data type and are generated either from a uniform or a Gaussian distribution. A regression model is trained on this random dataset to predict the target class. The predicted values are evaluated against the values of C^* to find the accuracy of the model obtained purely by chance, e.g. using the Pearson's correlation.

Definition 2: Classification from random: Let $D = \{C_1, ..., C_m\} \cup \{C^*\}$ be a random dataset of the shape $m \times n$ where C_1 to C_m are columns (features) in the dataset and C^* is the target column that partitions all n instances into r classes $Q_1 ... Q_r$. Let t_j be the true number of instances in each class Q_j. The categorical values of the features and the r classes are generated and assigned to the target randomly. A classification model is trained on the data set to predict the classes, which are then evaluated against the true classes (C^*) to find the accuracy of the classification model obtained purely by chance using some overlap metric, e.g F-Measure.

In this paper, the *shape* of a dataset is given by the number of features m, the number of samples n and the number of classes r: An $m \times n \times r$ dataset is a dataset consisting of n rows (each row referring to a data sample), m columns (each column referring to a feature) and r classes partitioning the samples into r partitions. We use the variable ρ to denote the ratio between the number of features and the number of samples, i.e. $\rho = m/n$. Furthermore, we use the term *wide dataset* to denote a dataset with $\rho > 10$.

IV. Chance influence on Prediction

In this section, we show that prediction accuracy measured for models trained using wide datasets can be for the most extent or even entirely caused by chance. We empirically quantify the extent of chance as a function of dataset shape. We also show that the way of performing feature selection as well as the validation modality, are key factors for avoiding chance influence. In particular, we generate random datasets according to Definitions 1 and 2. We train ML models (regression and classification separately), evaluate their performance and analyze the results in relation to the dataset shape and the modality of feature selection.

A. Dataset Generation

We generated three groups of random datasets, two consisting of numeric data and one consisting of categorical data. The first group (RDNUM1) contains 1000 datasets according to Definition 1, where the numbers of features as well as numbers of samples vary from 10 to 1000. The second group (RDNUM2) consists of 200 datasets according to Definition 1 with a fixed number of features, namely 10k, and sample sizes varying from 10 to 1000. This group is to represent very wide datasets like gene data. Feature values and target values in both groups are either drawn from a uniform distribution in the interval $[0, 1]$ or from a Gaussian distribution. The third group (RDCAT) consists of 1000 datasets according to Definition 2 with number of features as well as sample sizes varying from 10 to 1000. This group should represent categorical data. The feature values are either numerical like in RDNUM1 or random nominal values. The outcome is always categorical having the dummy nominal values C_0,C_1,.. C_r, where r is the number of classes, which varies from 2 to 9. All random values are generated using the random function implementation of Java 1.8.

B. Chance influence on Regression

The aim of this section is to demonstrate (i) that it is possible to train algorithms on pure random data and obtain high prediction accuracy due to chance correlation and (ii) that the way how feature selection is performed, strongly

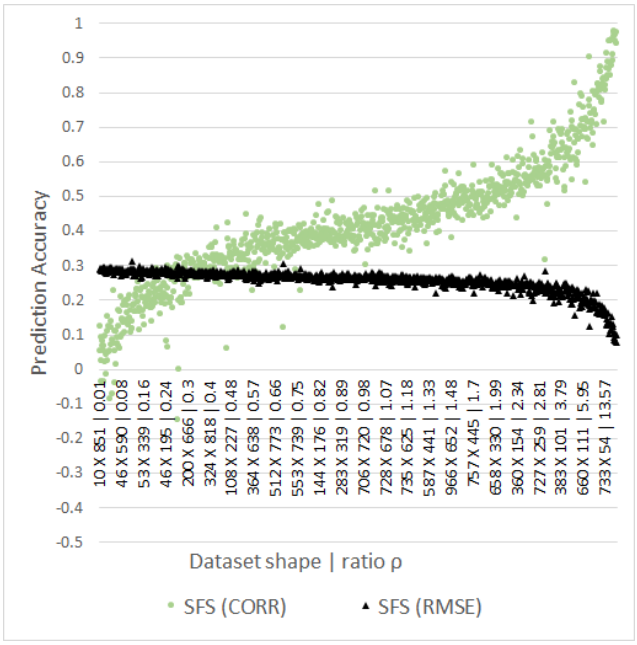

Fig. 1. Accuracy measures (Correlation and RMSE) of prediction Models trained on random datasets with different shapes using the single feature selection method (SFS). m and n vary from 10 to 1000 (Note that RMSE has here the same range $[0, 1]$ because the values in the target class are in the same interval).

affects the influence of chance on the prediction results. To this end, regression models have been trained on the datasets in the RDNUM1 group. As a feature selection, a simple BestFirst search and A Gaussian Processes classifier have been performed in a 10-fold cross validation. However, this feature selection has been performed in two different ways:

- *Single Feature Selection (SFS)*: The selection is performed only one single time prior to the cross-validation process using the whole dataset.
- *Folded Feature Selection (FFS)*: Individual feature selection for each split (fold) using the training data only (excluding the test part).

Figure 1 shows the correlation values (as accuracy measures) of 1000 prediction models, each trained on a dataset from the RDNUM1 group. It is clear that with single feature selection SFS, there is a strong increase of prediction correlation with increasing ρ (the ratio between the number of features and the number of samples). Almost perfect predictions are obtained when ρ is sufficiently high. Even when $\rho \approx 1$, there are still correlations in the order of 0.40. Figure 2 shows that the correlation with the modality Folded feature selection (FFS) are significantly lower. The vast majority of correlation values are between -0.1 and 0.1 and the RMSE values remains in the order of 0.3 which is the expectation value of RMSE for a random variable drawn from a uniform distribution in $[0, 1]$. Furthermore the accuracies have a significantly lower dependence on ρ. This is a clear indicator that the FFS feature selection modality mitigates the influence of chance and enable more reliable results.

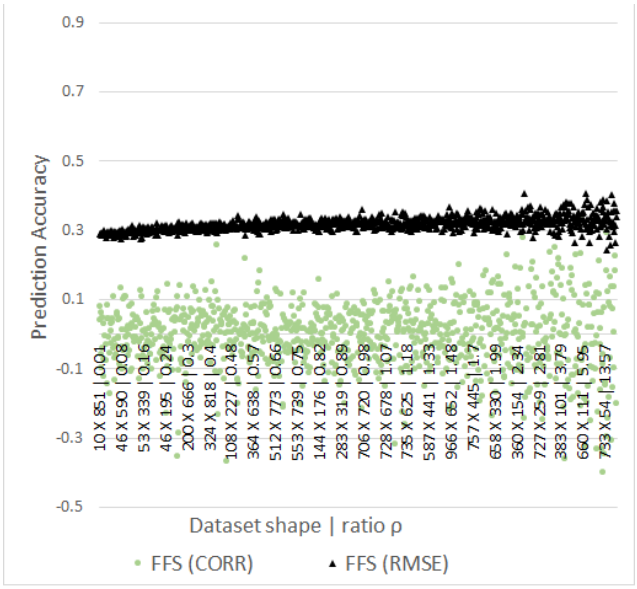

Fig. 2. Accuracy measures (Correlation and RMSE) of prediction Models trained on random datasets with different shapes using the folded feature selection method (FFS). m and n vary from 10 to 1000.

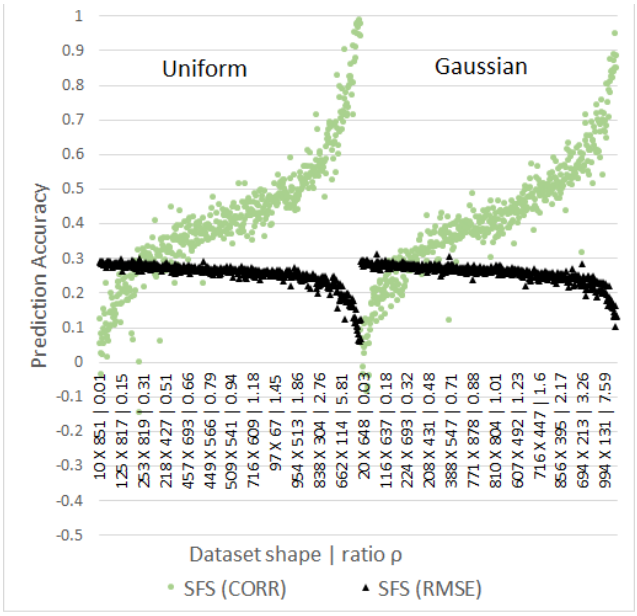

Fig. 3. Accuracy measures (Correlation and RMSE) of prediction Models trained on random datasets with different shapes using the single feature selection method (SFS) sorted according to distribution (uniform and Gaussian) and then according to ρ.

In Figure 3, the datasets are grouped by the data distribution and sorted by ρ to compare the uniform and Gaussian distributions. We see that the data distribution has almost no influence on the behavior of prediction accuracy from chance as a function of the dataset shape.

Figure 4 shows the chance behavior with very wide datasets, namely the datasets of the RDNUM2 group with 10000 features. It is remarkable to see that not only very high

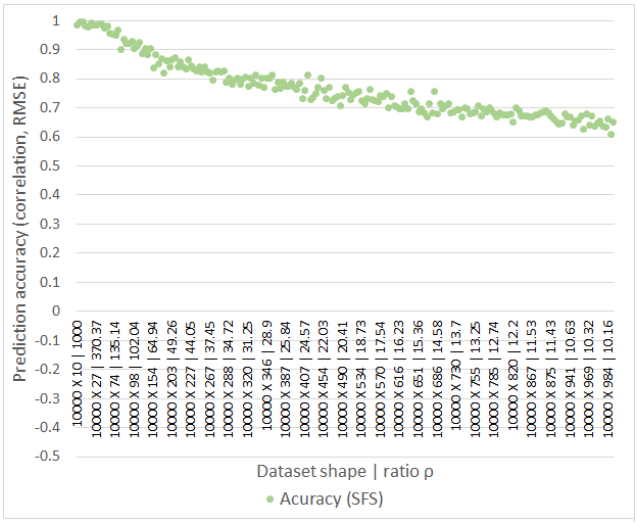

Fig. 4. Accuracy of prediction Models trained on random sets with 10k features and different sample sizes.

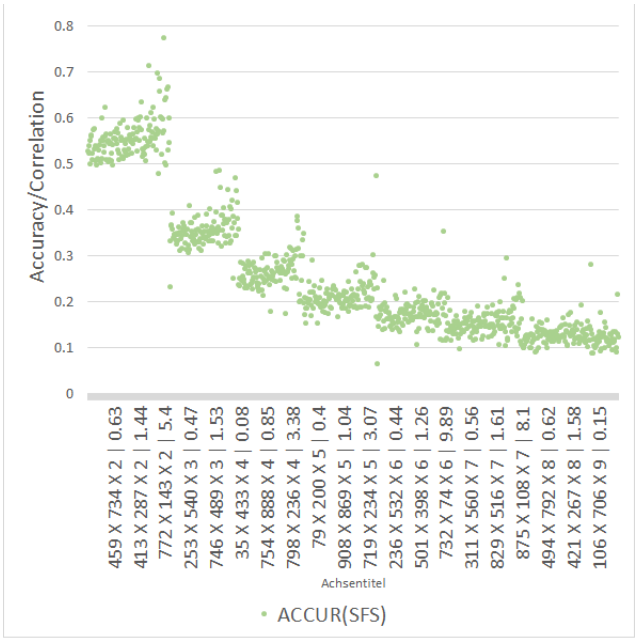

Fig. 5. Accuracy of prediction Models trained on random sets with different shapes. The datasets are first sorted according to ρ and then according to the number of classes.

accuracy can be obtained by chance with high ρ values, but that we still have accuracies around 0.5 even with sample sizes up to 1000 which significantly exceeds the typical sample size commonly used, e.g. in genomics research.

C. Chance influence on Classification

In this ssection we demonstrate the influence of chance correlation on classification, i.e. building classification model from random categorical (nominal) data. To demonstrate this, we trained J48 Trees on the 1000 datasets of RDCAT group using both SFS and FFS feature selection moralities. Figer 5 shows the accuracies of J48 trees trained trained on the datasets of the RDCAT group. The datasets are sorted at first according to ρ and then according to the number of target classes r. It is notable that the values are arranged in different levels. These levels relate to the number of target classes in the dataset.The classification accuracy strongly depend on these levels, i.e. on the number of classes r. There is already a dependence on ρ within each level, but this dependence decreases with increasing the number of classes r. It is interesting to see that there is almost no influence of ρ when there are more than 7 classes in the dataset.

V. Analysis

Observations and empirical results on random data show the considerable effect of chance on ML models. It is remarkable to see in Figure 4 that even using a sample size in the order of thousands, we still can have an additional accuracy caused by chance in the order of 60% to 70% according to state-of-the-art evaluation methods, given the number of features is such high as in the commonly used size in the genomics. In this section, we discuss the issue of chance influence from different view points in the light of the results of Section IV as well as the comments in the literature in this regards.

A. Feature Selection and chance

Feature selection methods are an essential step in ML that help exclude irrelevant features from the training process. This leads in many cases to performance enhancement, especially when the number of features is very large or when there is redundant information. Feature selection helps also as a dimensionality reduction to mitigate the effects stemming from the curse of dimensionality.

However, particularly in the case of large number of features, there are two types of features selected: (I) Features selected because they are indeed relevant, i.e. because they contain information with respect to the target class and (II) features selected because of their correlation by chance with the target class, i.e. they don't have information, but rather an accidental correlation in the underlying particular dataset. The probability of observing and selecting features of type II increases directly proportionally with the number of features and inversely proportionally with the number of instances.

Cross validation (CV) is a method to avoid over-fitting and increase generalizabilty. CV does not necessarily avoid chance influence, especially in its commonly used form, simply because the correlation stemming from chance in type II features technically does not differ from the correlation in type I features stemming from real information. However, depending on how cross validation is performed, chance influence can be considerably mitigated.

In Section IV-B, we presented two modalities of feature selection, namely the single feature selection (SFS) and the folded feature selection (FFS) and showed that the FFS considerably mitigates the chance influence, an observation that we will explain here:

- **SFS:** When a single feature selection is performed prior to a CV using the whole dataset, the feature selection step ends with a feature set that is reduced to those features that are relatively highly correlated with the target. Now splitting the dataset into folds does not change the fact that the remaining (selected) features are correlated in all instances and thus in all folds. This leads to the fact that a model trained on any nine folds will perform well on the tenth fold (assuming a 10-CV).
- **FFS:** In the folded feature selection, in each split another feature selection is performed using only the training subset. This leads to the following: Type I features selected based on the training subset will likely correlate also with the testing subset and thus lead to a higher score. On the contrary, Type II features will not correlate with the testing subset because they have accidental correlation with respect to the training subset only, thus lead to a lower score.

Of course, nothing comes without disadvantages: The result of performing a folded feature selection is different subsets of features, at least n subsets in an n-CV. This is not optimal if the aim is to identify the relevant features rather than to build a model.

B. Regression: Correlation versus RMSE

Taking a deeper look at Figure 1 and the data behind it, especially at the difference between the evaluation metrics correlation (CORR) and the root mean square error (RMSE), particularly their sensitivities to chance, one can note the following: While the CORR values span a range from 0 to 1, RMSE values remain between 0.33 and some values in the order of 0.1. The value 0.33 is the RMSE when the prediction goes totally wrong, i.e. random regarding the target class because this is the expectation value of the RMSE of uniformly distributed values in $[0, 1]$. It corresponds to zero correlation. On the oposit, the value RMSE$= 0$ corresponds to CORR$= 1$. Therefor, we normalize the RMSE values to be comparable with CORR values by $RMSE' = (0.33 - RMSE)/0.33$ to get the comparable plot in Figure 6. It shows that RMSE$'$ is in general less sensitive to chance than CORR: First it does not reach the extreme values (zero and one) like CORR and second RMSE$'$ is significantly less then CORR for the vast majority of the datasets. While CORR has a standard deviation σ^2 of 0.034, the RMSE hast a σ^2 of 0.015.

The observation above tells that the RMSE is preferable to be used as a quality measure instead of CORR, when the data is suspected to affected by chance, e.g. with a large number of features and/or a small sample size.

C. Classification: Number of classes r

Figure 5 showed the accuracies of classification models trained purely by random data. The values are clustered in different groups, where each group refers to a value of r (the number of classes), which results in different levels of accuracy. The average accuracy in each cluster (level) strongly depends on r. Actually, the minimum value in each cluster is equal to $\frac{1}{r}$, which is the probability of true assignment of an object by random, given r classes (assuming equal distribution of the objects to the classes). The increase of accuracy above this value stems from ρ, that is having additional features results in an increase of the accuracy above $\frac{1}{r}$. However, this increase is not the same in all levels, but rather decreases with increasing r, i.e. the more classes there are, the less influence ρ has on the accuracy. The influence of dimensionality ρ almost vanishes when r is more than seven. We can conclude that classification models become significantly less prone to chance influence with increasing number of classes.

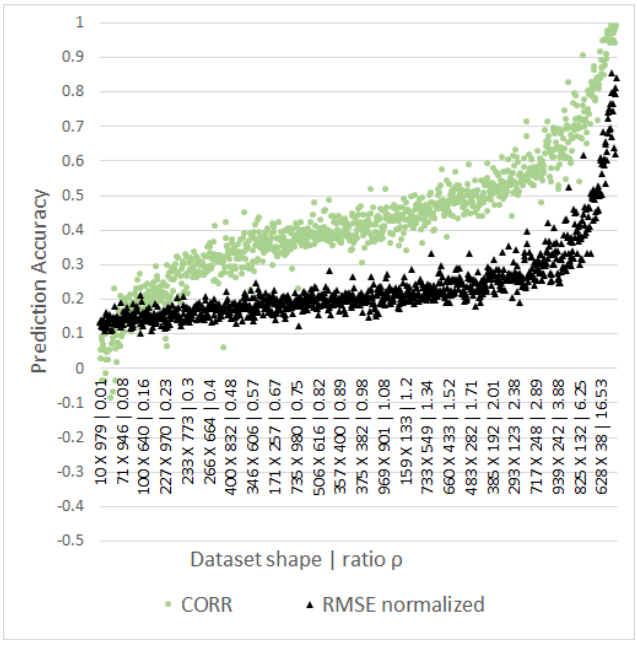

Fig. 6. RMSE normalized to be comparable with CORR. Quality measures of prediction Models trained on random datasets with different shapes using single feature selection (SFS) method. m and n vary from 10 to 1000.

D. Correction for chance

Evaluation metrics that correct for chance are not a new thing. The Cohen's Kappa metric [12] for example calculates the agreement between two raters, thereby considering the chance agreement, i.e. it corrects the agreement down based on the expected chance. The Kappa is defined as

$$Kappa = \frac{A_0 - A_e}{1 - A_e} \tag{1}$$

where A_0 is the agreement (e.g. overlap) between two rater (in our the true classification and the prediction) and A_e is the hypothetical probability of chance agreement. For a classification with n objects and r categories, A_e is defined as:

$$A_e = \frac{1}{n^2} \sum_{i=1}^{r} N_{1i} N_{2i} \tag{2}$$

where N_{ji} the number of objects predicted by rater j as Class i. For example, assume that you have n objects, where a half of them is assigned to Class 1 and the other half to Class 2. Assume that there is a dummy classifier that assigns

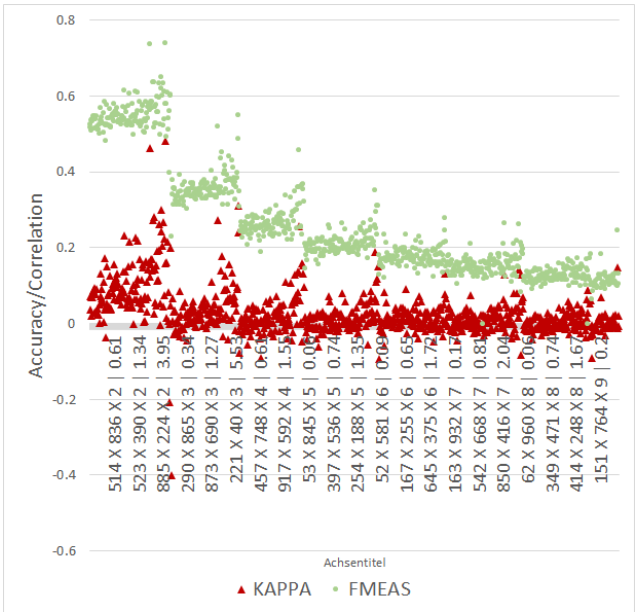

Fig. 7. Cohen's Kappa and F-measure as evaluation of prediction Models trained on the RDCAT datasets. Kappa measure corrects the values for chance by shifting them down but does not correct the accuracy increase stemming from ρ.

the classes randomly and you apply both of the F-Measure and the Kappa metrics to evaluate its accuracy. While the F-measure will score the dummy classifier with 0.5, the Kappa will score it with zero.

Equation 2 tells that Kappa considers the number of the classes as well as how the objects are assigned to the classes. It does not consider the factors that lead to these assignment. In other words, it calculates the hypothetical probability of chance just based on the object-class distribution and does not consider the number of features/instances used to find out this distribution. To demonstrate this fact, we evaluated the datasets of the RDCAT group additionally using the Kappa measure and plotted them beside the F-measure in Figure 7. If Kappa measure were able to completely correct the chance, all kappa values would be zero as expected. Kappa measure is shifted down but still has accuracy that is not corrected, namely the accuracy stemming from ρ.

This observation motivates defining a new correction for chance that additionally takes into account the number of features in relation to the number of instances under consideration of the number of classes, which is one topic of our future work.

Conclusion

We showed that in datasets with a very large number of features, like genomics datasets, chance is so considerable that it can be responsible for very high accuracies of classification and regression models. If ignored, chance could be a factor that leads to accurate, but not generalizable models. We showed that the way how feature selection is performed has a significant impact on chance influence despite cross validation, a fact that justifies to recommend using the folded feature selection within cross-validation. We also showed that the tendency of classification to be influenced by chance significantly decreases with increasing number of classes. We finally showed that different evaluation metrics are differently prone to chance. However, even the kappa metric, which is designed to correct for chance, cannot correct the chance stemming from dataset dimensionality in the sense of the ratio between number of features and sample size. These facts motivate us to continue this research to (i) formally estimate the chance in a dataset based on the dataset dimensionality expressed by the numbers of instances, features, and classes (ii) test and compare other metrics regarding their proneness to chance, (iii) extend metrics like the Kappa to consider the chance stemming from dataset dimensionality and (iv) investigate the settings that minimize the influence of chance on training and evaluation.

References

[1] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proceedings of the National Academy of Sciences*, vol. 103, no. 15, pp. 5923–5928, 2006. [Online]. Available: http://www.pnas.org/content/103/15/5923

[2] X. Fan, L. Shi, H. Fang, Y. Cheng, R. Perkins, and W. Tong, "Dna microarrays are predictive of cancer prognosis: A re-evaluation," *Clinical Cancer Research*, vol. 16, no. 2, pp. 629–636, 2010. [Online]. Available: http://clincancerres.aacrjournals.org/content/16/2/629

[3] J. P. A. Ioannidis, "Why most published research findings are false," *PLOS Medicine*, vol. 2, no. 8, 08 2005. [Online]. Available: https://doi.org/10.1371/journal.pmed.0020124

[4] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *The Lancet*, vol. 365, no. 9458, pp. 488 – 492, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0140673605178660

[5] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty, "Optimal number of features as a function of sample size for various classification rules," *Bioinformatics*, vol. 21, no. 8, pp. 1509–1515, 2005. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bti171

[6] K. Fukunaga and R. R. Hayes, "Effects of sample size in classifier design," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 8, pp. 873–885, Aug 1989.

[7] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo, "Predicting sample size required for classification performance," *BMC Medical Informatics and Decision Making*, vol. 12, no. 1, p. 8, Feb 2012. [Online]. Available: https://doi.org/10.1186/1472-6947-12-8

[8] S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T. R. Golub, and J. P. Mesirov, "Estimating dataset size requirements for classifying dna microarray data," *Journal of Computational Biology*, vol. 10, pp. 119–142, 2003.

[9] K. K. Dobbin, Y. Zhao, and R. M. Simon, "How large a training set is needed to develop a classifier for microarray data?" *Clinical Cancer Research*, vol. 14, no. 1, pp. 108–114, 2008. [Online]. Available: http://clincancerres.aacrjournals.org/content/14/1/108

[10] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proceedings of the National Academy of Sciences*, vol. 99, no. 10, pp. 6562–6566, 2002. [Online]. Available: http://www.pnas.org/content/99/10/6562

[11] J. Kuligowski, D. Perez-Guaita, J. Escobar, M. Guardia, M. Vento, A. Ferrer, and G. Quintas, "Evaluation of the effect of chance correlations on variable selection using partial least squares-discriminant analysis," vol. 116, pp. 835–40, 11 2013.

[12] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, p. 37, 1960.