ComDia+: An Interactive Visual Analytics System for Comparing, Diagnosing, and Improving Multiclass Classifiers

Chanhee Park Jina Lee Hyunwoo Han Kyungwon Lee

Ajou University

ABSTRACT

Performance analysis is essential for improving classification models. However, existing performance analysis tools do not provide actionable insights such as the cause of misclassification. Machine learning practitioners face difficulties such as prioritizing model, looking over confusion between classes. In addition, existing performance analysis tools that provide feature-level analysis are difficult to apply to image classification problems. This study has been proposed to solve these difficulties. In this paper, we present an interactive visual analytics system for diagnosing the performance of multiclass classification models. Our system is able to compare multiple models, find weaknesses, and obtain actionable insights for improving models. Our visualization consists of three views for analyzing performance at the class, confusion, and instance levels. We demonstrate our system using MNIST handwritten digits data.

Keywords: Performance analysis, classification, model comparison, confusion analysis.

Index Terms: H.5.2 User Interfaces: Graphical user interfaces.

1 INTRODUCTION

Classification is a common task in machine learning for modeling problems that have more than two disjoint classes [10, 11]. Performance analysis is essential for improving a classification model. Trough performance analysis, practitioners select an appropriate classifier, tune parameters, collect additional data, refine data, and extend to the classifier in order to incorporate domain knowledge or to handle special cases. A common approach to understanding model performance in machine learning is to look at performance measures such as accuracy, precision, and recall. These measures provide a summary of model performance.

However, in multiclass classification, analysis with these measures is difficult if you have many models and classes to compare. For example, consider a comparison of two models with the same overall accuracy. One model shows similar performance in all classes, while the other model shows very high performance in certain classes and poor performance in other classes. The overall performance measures of the two models are identical, but the evaluation of the two models should be different.

The most commonly used visualization for classification performance analysis is the confusion matrix [2]. It shows the misclassification behavior of a classifier by placing predicted and actual values in each row and column. Compared with overall performance measures, this matrix provides more details about the results and helps with introducing appropriate adjustments to the classifier. However, it does not provide solid evidence of the cause of the misclassification and how to improve it. Therefore, classifier development and debugging take longer and are more prone to misclassification.

Recent performance analysis visualization systems show the feature values of misclassified instances to find evidence of the cause of misclassification and how to improve the model. However, feature analysis is not effective when the classification problem takes with data that cannot be analyzed at the feature level. In particular, many recent studies have focused on image classification problems. In image classification problem, each feature is a pixel which human cannot analyze. Therefore, a performance analysis system for image classification models requires a new method that does not depend on features.

This paper presents a visualization system that effectively compares performances of multiple multiclass classifiers in image classification problem. This system shows the misclassification patterns of each classifier and provides actionable insights such as the causes of misclassification to improve performance. With this system, practitioners can compare the performance of models, diagnose weaknesses in each model, and find ways to address these weaknesses. Also, practitioners can analyze the performance of the model overall, class-wise, or instance-wise. The main contributions of this paper are as follows:

- Compare: Design of visualization to compare the performance of two or more classifiers by classes.
- Diagnosis: Design of visualization that allows the classifier to identify the cause of the misclassification in image classification problem.
- Improve: Design of visualization to know how to correct misclassifications of the classifier or improve overall performance in image classification problem.

2 RELATED WORKS

Various approaches have been proposed to improve classification performance using visualization. Previous studies have focused on understanding predicted results from classifiers or improving the classifier performance interactively.

Several studies have suggested effective model performance visualization to identify the overall model performance and to diagnose weaknesses. Amershi et al. propose ModelTracker [6], which can debug model performance issues by projecting the prediction results in a 2D space. This visualization cannot be applied to multiclass classification or multiple model comparisons. Ren et al. present Squares [5], which is an instance-based performance visualization for multiclass classification problems. The advantage of this visualization is that it uses less space and has faster identification performance than the confusion matrix. However, it does not provide the causes of misclassifications or insights to improve performance. Rule Matrix [7] understands the model by extracting the rules from the behavior of the model. It extracts a standardized rule-based knowledge representation from the model's input-output behavior. It makes it possible for machine

* Chanhee Park, Jina Lee, Hyunwoo Han, and Kyungwon Lee are with the Ajou University. E-mail: {ch13p, qqwwdj, ainatsumi, kwlee} @ajou.ac.kr

learning non-experts to understand the behavior of the model, but the limitation is that the rules do not completely match the model.

Other studies refine data or focus on finding effective ensemble cases to improve model performance. Talbot et al. [8] propose Ensemble Matrix, which consists of the small multiples of confusion matrices. It shows several combinations of two classifiers for making a more effective model ensemble. Zhao et al. propose LoVis [9], which finds the optimal data segment point to improve the performance of the linear model.

Some recent studies propose systems that can effectively diagnose and improve model performance by providing a feature-level or instance-level analysis. Alsallakh et al. propose Confusion Wheel [3], which analyzes probability classifications using radial layout. It shows the number of instances of true/false or positive/negative according to the predicted confidence by class. This study has limitations in comparing the superiority of performance over multiple models. Zhang et al. present Manifold [4], which is a model-agnostic framework for the interpretation and diagnosis of machine learning models. It provides two views for comparison of model performance and analysis of features so practitioners can explore the cause of misclassification. It has the advantage of considering complementary models through performance comparison of model pairs. However, it is difficult to compare the performance of three or more models at once and to apply to data (e.g., images) that cannot be analyzed in terms of features.

3 Motivation

We designed the visualization goals based on previous work investigating current practices and difficulties faced by machine learning practitioners encounter when creating multiclass classifiers [5]. Users consider model prioritization as important but the most difficult task. Therefore, when comparing multiple models, it is necessary to provide not only measures of model performance, but also their rank. Practitioners regard understanding the overall performance of a model as relatively easy. However, they find it difficult to confuse classes but feel it is an important task. Therefore, visualization is needed to help prioritize model improvements at the level of confusion between classes.

Many previous studies have provided feature-level analysis to find out the cause of the model's misclassification [2, 4]. However, as the number of features increases, it is difficult to infer the cause of the misclassification and to improve the model through feature analysis. In particular, many recent machine learning studies have focused on image classification problems. [17, 18] In these studies, pixels in each image are used as features for classification. However, humans cannot conduct the pixel-level analysis. Therefore, a new visualization method is necessary to find the cause of error in image classification.

Some study noted that misclassification patterns between models are different [3, 4]. Since each model has different instances that cause misclassification, it is possible to improve misclassification through a proper combination of models. Therefore, there is a need for a visualization design that can find an appropriate combination of models to improve model performances.

4 Visual analysis system

Our visualization shows classification results of one datum by multiple classifiers. In this section, we describe three major design goals. We then describe a dataset and five classifiers used to illustrate our visualization. Finally, we propose our visualization system to analyze multiple classifiers.

4.1 Design Goals

The goals of this study are as follows:

G1: Design visualization that compares and ranks performances of multiple models at the class level for giving insight into model selection.

G2: Design visualization that shows performance of model and cause of misclassification based on confusion between classes in image classification problem.

G3: Design visualization that gives insight into improved classification performance.

4.2 Data Set and Classifiers

We illustrate our system using the MNIST database [1]. The MNIST database is a large database of handwritten digits that is commonly used for training and testing in the field of machine learning. We used this dataset for the following reasons. First, it is made up of images. Secondly, it has many classes (10 classes: handwritten digits from 0 to 9). Finally, it is easy to understand because it is familiar to practitioners.

We have made five classifiers for visualization. If the performance gap between the classifiers is large, in-depth analysis is not required. Therefore, we limited the performance difference to within 4% (90% ± 2%). The name and accuracy of each classifier are as follows:

Name of the Classifier	Accuracy
M-0: Stacked Auto Encoder	0.9177
M-1: Neural Network (3 layer)	0.9174
M-2: Random Forest	0.9038
M-3: Softmax Regression	0.8912
M-4: kNN (k is 3)	0.8843

Table 1: The name and accuracy of the classifiers.

4.3 Visualization and Interaction

Our visualization has three views. 1) A ranking view for comparing the performance of classifiers, 2) a matrix view for diagnosing the weaknesses of a classifier, and 3) a correct misclassification view for gaining insight on how to improve performance. To ensure enough visibility of the information, up to 20 classes and 10 models can be visualized. A previous study has confirmed that most multiclassification problems cover less than 20 classes [5]. Therefore, our visualization is applicable to these problems. And some performance analysis studies do not provide two model comparisons or provide only 2 model comparisons [3, 4, 5, 6]. Sometimes, however, a machine-learning practitioner must compare three or more models. Our visualization system comparing up to 10 models has strengths compared to previous studies.

4.3.1 Performance Ranking View

This view shows the performance measures and ranks of multiple models at the class level (G1). Measures/ranks mean absolute/relative performance. This view shows the superior and inferior class in each model's performance. From this view, the user can select a final model and find out the model weaknesses (G1).

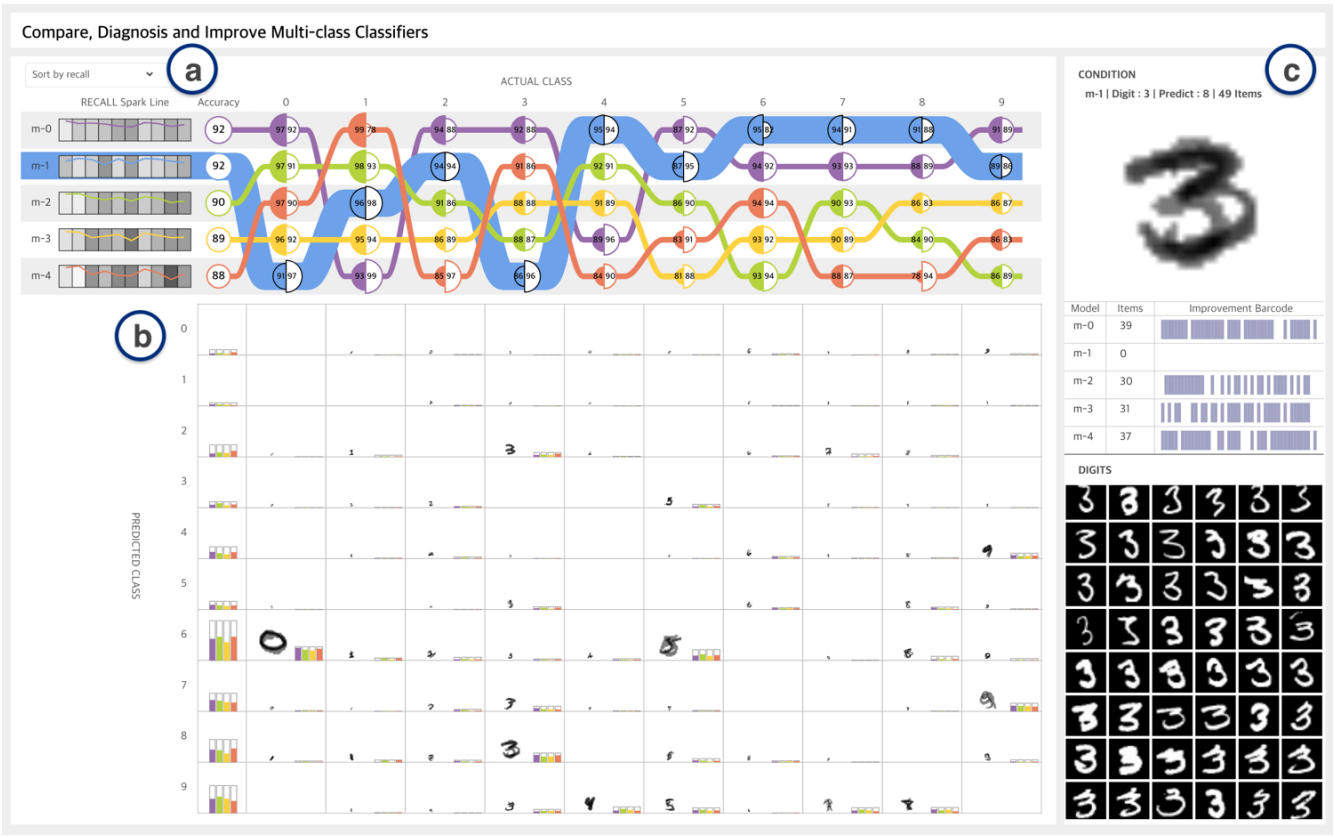

Figure 1: The interactive exploration environment shows information about performance, weakness, and insights for improvement. (a) compares the overall performance of classifiers at the class level. (b) shows the weakness of a classifier at the confusion level. Each cell has an instance composite image that is proportional to the size of the misclassification and a bar chart that shows the amount of prediction improvement through different models. (c) provides insight into performance improvements by showing misclassification instances and comparing the predictions with other models.

In this view, each column represents a class, and each row represents the model. The performance of each model-class combination is represented by the size of a circle. The left half of each circle is recall and the right half is precision. This model-class performance circle is sortable according to four user options: model name, overall accuracy, recall value of each class, and precision value of each class. When a model is sorted by name or overall accuracy, one model is in one row. However, when the model is sorted by recall or precision, the model corresponding to each row changes for each class. Each model has a representative color. The model-class performance circle is connected by a color line. When the user hovers their mouse on or selects a model, this line becomes thicker. The weakness of the model can be found by comparing the thickness and the circle size. Because the thickness corresponds to the diameter of the circle when recall/precision is the average performance of the overall model, if the size of the circle is less than the thickness of the line, it is a weakness.

It is difficult to know the performance trend of one model when many models are shown in ranking. Sparklines and heat maps provide a summary of each model's recalls or precisions. Sparklines represent the recall/precision per class of a model in line height. This shows whether model performance is stable or has weaknesses in certain classes. Sparklines are suitable for determining the performance within a model but are not suitable for comparison between different classes. A performance heat map provides a recall/precision comparison between different models and classes. We use a metaphor that the misclassification is models' impurity. Therefore, the lower the performance, the darker the color. This shows whether the performance between the other model and the other class is higher or lower.

4.3.2 Diagnosis Matrix View

The user can select a model in the performance ranking view. The diagnosis matrix view displays the diagnosis matrix of the selected model. This view is based on the confusion matrix. Columns are actual classes, and rows are predicted classes. Each cell shows how many instances have made misclassifications at the confusion level (combination of actual and predicted class is incorrect) (G2). Our aim was to analyze and improve misclassification at the confusion level through this matrix view. Therefore, we introduced two improvements to the confusion matrix.

Figure 2: Comparison of two average images where a neural network incorrectly predicts digit 3 as 7 and 3 as 8. Users use this information when refining and collecting more data

First, the diagnosis matrix shows the cause of the misclassification at the confusion level for the image classification problem (G2). Each cell in the matrix shows an average image of the instances corresponding misclassification. For example, in row 4, line 9, there is an average image of instances that incorrectly predicted the number 4 as 9 since columns are actual classes and rows are predicted classes. The larger the number of misclassifications, the larger the image. This shows not only the weakness of the model but also the cause of the misclassification. The user can refine or collect more data using this information (G3).

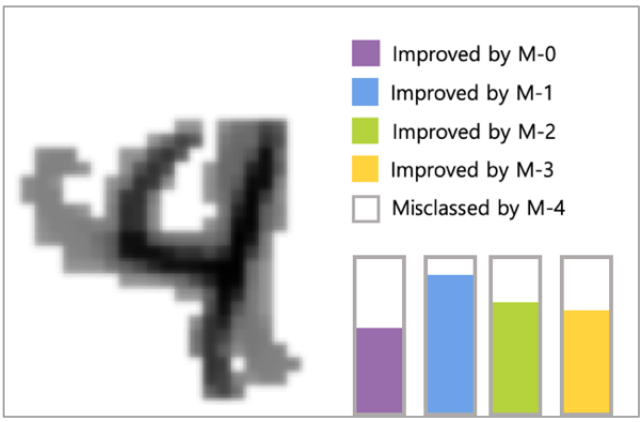

Figure 3: Improvement bar chart on misclassifications where kNN (M-4) incorrectly predicts the digit 4 as a 9.

Secondly, the diagnosis matrix suggests improvements through comparison with other models. Each cell has a bar chart. This bar chart shows how correctly the other models predict the wrong prediction in the selected model. In other words, the bar height represents the number of instances that can be improved by other models. This shows the possibility of improvement by comparing. with other models. Users integrate models or add specific extensions through this information (G3).

4.3.3 Correct Misclassification Instance View

The user can select any misclassification case (each actual predict combination cell) in the diagnosis matrix view. The correct misclassification view shows specific information about the selected misclassification case. This view is displayed after model selection and misclassification selection. At the top of this view, the selected model name, actual class, predicted class, and number of misclassification instances are displayed so that users remember. This also shows largely the average image of the misclassification instances according to the selected condition.

The bar chart in the diagnosis matrix view shows how correctly the other models predict the wrong prediction in the selected model. However, it does not show how much the model-to-model improvement instances overlap. Therefore, the bar chart is inappropriate when the user wants to refer to two different models to improve the model. We solve this problem through rug plots. A rug plot, also known as a 1-dimension scatter plot, displays instances as marks along an axis. This represents the improved instance in one line. Many lines and high density in the rug plot mean that many instances are improved by the model. The rug plot shows the distribution as well as the number of instances. If high-density points in the rug plot are different, the two models will improve on different instances. The user can sort the instances in rug plots according to whether a specific model corrects such instances. This interaction makes it easier to compare instance distributions between models(G3).

Also, this view shows all instance images that satisfy the selected condition. Sometimes there is less similarity between images with misclassifications, or one or two unique images can be very different from other images. In such a case, the average image of the incorrectly predicted instance is difficult to interpret. Therefore, our visualization shows the original image of every instance at the bottom of this view.

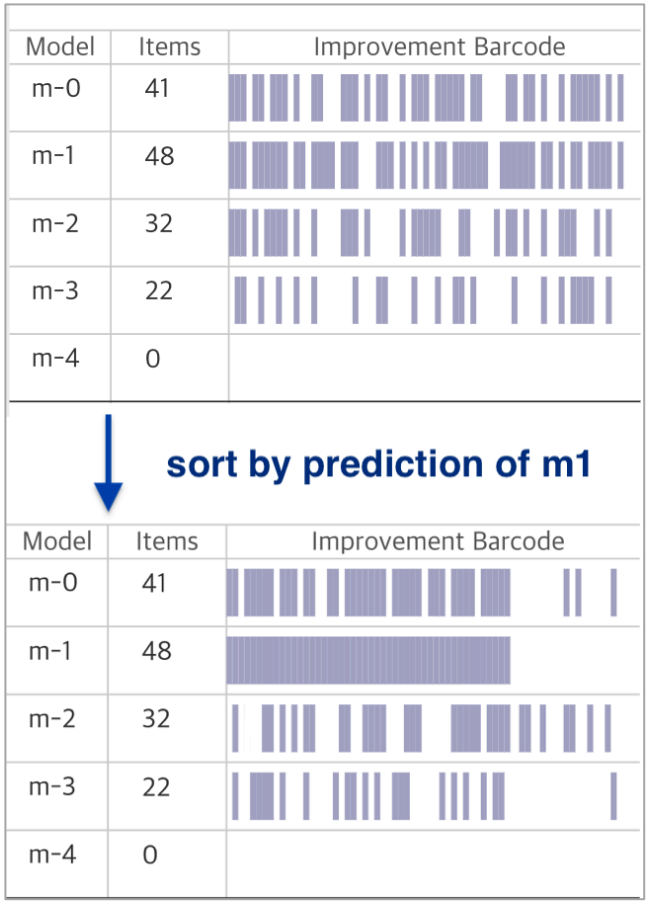

Figure 4: Improvement barcode chart on misclassifications where kNN (M-4) incorrectly predicts the digit 7 as a 1. M-1 is the best for improving incorrectly predicted instances. M-2 improves incorrectly predicted instances where M-1 cannot improve.

5 Conclusion

We presented an interactive performance visualization for comparing, diagnosing, and improving multiclass classifiers. This visualization compares the performance of multiple classifiers at the class level. Also, it provides information to improve performance at the confusion level. We also proposed an instance level analysis method for image classification to find the cause of misclassification.

In our visualization, information related to one model is scattered in the three views, and the average image technique is hard to apply if images are complex. Our future work will solve these problems. We will assemble information related to one model in one view. Also, we will introduce a method to extract common features from multiple images so that they can be applied to complex image classification tasks. Our future work will include evaluations to verify that the proposed visualization is actually helpful to practitioners.

Acknowledgments

We thank Oh-Hyun Kwon at the University of California, Davis for his support and valuable feedback.

REFERENCES

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE*, 86(11):2278-2324, November 1998.

[2] Townsend, James T. “Theoretical analysis of an alphabetic confusion matrix”. Perception & Psychophysics, 9(1):40–50, 1971.

[3] Alsallakh, Bilal, et al. "Visual methods for analyzing probabilistic classification data." IEEE transactions on visualization and computer graphics 20.12 : 1703-1712, 2014

[4] Zhang, Jiawei, et al. "Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models." IEEE transactions on visualization and computer graphics, 2018.

[5] Ren, Donghao, et al. "Squares: Supporting interactive performance analysis for multiclass classifiers." IEEE transactions on visualization and computer graphics 23.1 : 61-70, 2017.

[6] Amershi, Saleema, et al. "Modeltracker: Redesigning performance analysis tools for machine learning." Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM, 2015.

[7] Ming, Yao, Huamin Qu, and Enrico Bertini. "RuleMatrix: Visualizing and Understanding Classifiers with Rules." IEEE transactions on visualization and computer graphics. 2018.

[8] Talbot, Justin, et al. "EnsembleMatrix: interactive visualization to support machine learning with multiple classifiers." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2009.

[9] Zhao, Kaiyu, et al. "Lovis: Local pattern visualization for model refinement." Computer Graphics Forum. Vol. 33. No. 3. 2014.

[10] Cireşan, Dan, Ueli Meier, and Jürgen Schmidhuber. "Multi-column deep neural networks for image classification." *arXiv preprint arXiv:1202.2745,* 2012.

[11] Golub, Todd R., et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." science 286.5439, 531-537, 1999.

[12] Kwon, Bum Chul, et al. "Clustervision: Visual supervision of unsupervised clustering." *IEEE transactions on visualization and computer graphics* 24.1 :142-151 2018.

[13] Schneider, Bruno, et al. "Visual Integration of Data and Model Space in Ensemble Learning." *arXiv preprint arXiv:1710.07322*, 2017.

[14] Tam, Gary KL, Vivek Kothari, and Min Chen. "An analysis of machine-and human-analytics in classification." *IEEE transactions on visualization and computer graphics* 23.1 :71-80, 2017.

[15] Gratzl, Samuel, et al. "Lineup: Visual analysis of multi-attribute rankings." *IEEE transactions on visualization and computer graphics* 19.12 :2277-2286, 2013.

[16] Stolte, Chris, Diane Tang, and Pat Hanrahan. "Polaris: A system for query, analysis, and visualization of multidimensional relational databases." *IEEE Transactions on Visualization and Computer Graphics* 8.1 : 52-65, 2002.

[17] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[18] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.