EuroVis Workshop on Visual Analytics (2015)
E. Bertini and J. C. Roberts (Editors)

Interactive Image Feature Selection Aided by Dimensionality Reduction

P. E. Rauber[1,2], R. R. O. da Silva [1,3], S. Feringa[1], M. E. Celebi[4], A. X. Falcão [2], A. C. Telea[1]

[1] Johann Bernoulli Institute, University of Groningen, the Netherlands
[2] Institute of Computing, State University of Campinas, Brazil
[3] Institute of Mathematical and Computer Sciences, University of São Paulo, Brazil
[4] Department of Computer Science, Louisiana State University in Shreveport, United States of America

Abstract

Feature selection is an important step in designing image classification systems. While many automatic feature selection methods exist, most of them are opaque to their users. We consider that users should be able to gain insight into how observations behave in the feature space, since this may allow the design of better features and the incorporation of domain knowledge. For this purpose, we propose a methodology for interactive and iterative selection of image features aided by dimensionality reduction plots and complementary exploration tools. We evaluate our proposal on the problem of feature selection for skin lesion image classification.

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [INFORMATION INTERFACES AND PRESENTATION]: User Interfaces—User-centered design

1. Introduction

Image classification is a widely studied problem in image analysis and computer vision. Most classification systems can be divided into two steps. Firstly, *feature extraction* represents each input image by a high-dimensional feature vector. Next, these vectors are *classified* by support vector machines, random forests, artificial neural networks or other supervised learning techniques [ZIL12,CS13].

Between feature extraction and classification, an extra step can occur: *feature selection*. In this step, a subset of features is selected to either improve the accuracy of subsequent prediction (classification or regression) tasks, increase computational efficiency, or enhance the understanding of the data [GE03].

Classification systems are typically opaque: they do not explain to their users how they arrive at decisions. When such systems perform incorrectly for particular input data, understanding how to improve them may become a daunting task. Even when they perform correctly, users often prefer systems that can explain their decisions in understandable terms [KG12]. As an example, the study in [DB05] showed that dermatologists consider that an ideal diagnosis system should explicitly justify its decisions and define its confidence level. Naturally, a smaller number of features leads to easier inspection and interpretation. At the same time, significant effort is required to design effective sets of features, and the computational cost of computing features may need to be factored into decisions. For these reasons, feature selection is a challenging and important topic in image analysis.

We propose a visual analytics approach to help classification system designers with feature selection tasks. Our approach, and associated tool, supports an iterative, incremental workflow where users have immediate visual feedback on their choices via a 2D projection of observations restricted to the features under inspection. We propose views that allow reasoning about feature subspaces, and integrate them with feature scoring techniques that help bootstrapping the feature selection process. We demonstrate our approach in the task of feature selection for skin lesion image classification.

Section 2 reviews related work. Section 3 presents our proposal and supporting tool. Section 4 demonstrates the tool in the context of feature selection for skin lesion image classification. Section 5 details the implementation of the tool. Section 6 discusses our proposal. Section 7 concludes the paper.

2. Related work

Related work covers the tasks of feature extraction, feature selection, and feature space exploration, as follows.

Feature extraction can be performed on entire images or presegmented regions of interest. Typical features are related to color, texture, shape, and spatial characteristics of image elements [ZIL12]. Another popular image description method is bag-of-visual-words, which uses a histogram over a predefined list of image patches [Tsa12]. More recently, classification systems based on deep networks [KSH12, Ben09] became able to deal directly with raw image data, bypassing feature extraction. However, these networks usually require large amounts of data for training.

Feature selection has been widely studied in machine learning [GE03]. Numerous heuristic methods have been proposed for this task [LY05]: *filters* rely on data characteristics without involving prediction algorithms; *wrappers* base their selection on a prediction algorithm; and *hybrids* combine these two approaches. Beyond selecting a feature subset, some of these methods base their decisions on numerically scoring (or ranking) features.

Feature space exploration may involve the visual search for structures and patterns in high-dimensional data spaces. Classical techniques for this purpose include scatterplot matrices and parallel coordinates [BTK11]. Another class of techniques apply dimensionality reduction. This process finds a low dimensional representation of the data that retains its *structure*, which is defined by relations between points, presence of clusters, or overall spatial data distribution [LWBP14]. Numerous techniques address the interactive exploration of high-dimensional feature spaces. Tatu *et al.* propose finding interesting feature subspaces and displaying the data in these subspaces; however, the methods employed do not scale well to hundreds of dimensions [TMF*12]. Krause *et al.* aid feature selection via the visualization of aggregated feature relevance data [KPB14]. In contrast to our work, they do not provide an integrated representation of the feature space. Turkay *et al.* propose exploration using 2D representations of both observations and features [TFH11].

Most related to our work, Yuan *et al.* present a tool that, among other functionalities, displays 2D projection plots of observations restricted to selected feature subsets [YRWG13]. Their goal was to allow subspace cluster exploration, while we focus on aiding feature selection for classification tasks.

3. Proposed methodology and tool

An observation is a vector $\mathbf{x} \in \mathbb{R}^m$ that describes an object of interest, which is an image in our case. The j-th element $x_j \in \mathbb{R}$ of $\mathbf{x}$ is also called feature j. A feature corresponds to a quantity that is measured directly from an object (e.g., average luminance). We denote the set of all n observations under study by $\mathcal{X} = \{\mathbf{x}^{(i)} | 1 \leq i \leq n\}$, and the set of all features by $\mathcal{F} = \{1, \ldots, m\}$. For any $\mathcal{F}' \subseteq \mathcal{F}$, having $d \leq m$ features, we denote by $\mathbf{x}_{\mathcal{F}'} \in \mathbb{R}^d$ the observation $\mathbf{x}$ restricted to features in $\mathcal{F}'$. We let $\mathcal{X}_{\mathcal{F}'}$ denote the set $\mathcal{X}$ restricted to features in $\mathcal{F}'$. In supervised learning, *classification* assigns a class $c \in \mathcal{C}$ to each observation $\mathbf{x} \in \mathcal{X}$ based on previous experience. Feature selection aims to find a feature subset $\mathcal{F}' \subseteq \mathcal{F}$ that is small and sufficient for generalization given $\mathcal{X}_{\mathcal{F}'}$.

Let a k-dimensional projection be a function $P : \mathbb{R}^m \to \mathbb{R}^k$, where m is usually large (hundreds) and k is usually 2. If P preserves the *structure* (as already defined) of the observation set $\mathcal{X}$, we can use the projection $\mathcal{X}_P$ to reason about $\mathcal{X}$ [LWBP14]. This is useful because $\mathcal{X}_P$ can be represented visually (e.g., as a scatterplot).

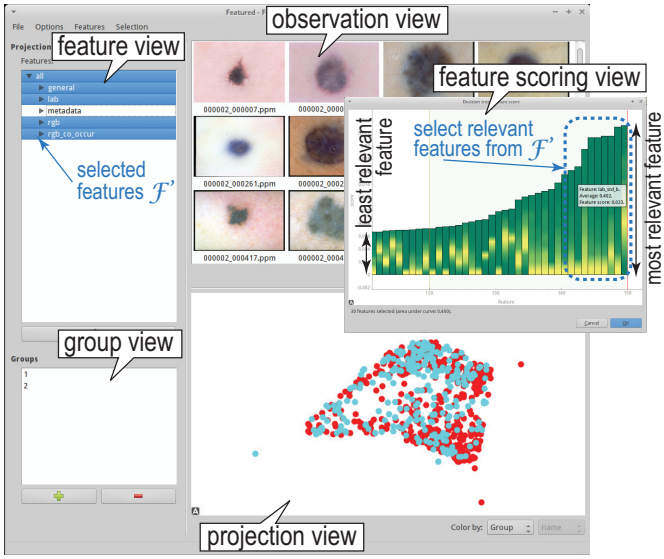

Figure 1: *Tool overview*

In this work, we propose feature selection guided by interactive and iterative 2D projections of high-dimensional feature spaces. For this purpose, we developed an interactive tool composed of five linked views (Fig. 1). The *observation view* displays the image associated to each observation $\mathbf{x} \in \mathcal{X}$, optionally sorted by a feature of choice. This provides an easy way to verify if a feature corresponds to user expectations. The *feature view* shows all features $\mathcal{F}$, optionally organized as a hierarchy based on semantic relations. For example, color-related features may be grouped into a single node. Within this view, users can select a feature subset $\mathcal{F}'$ to further explore. The *group view* allows the creation and management of observation groups by direct selection in the observation view or in the projection view (presented next). The *projection view* shows a scatterplot of the 2D projection of $\mathcal{X}_{\mathcal{F}'}$, the set of all observations restricted to the selected feature subset $\mathcal{F}'$. Plot points can be colored by a user-selected feature or user-defined groups, and are highlighted to show the selected set of observations $\mathcal{X}' \subseteq \mathcal{X}$. Finally, the *feature scoring view* ranks the features in $\mathcal{F}'$ sorted by a relevance metric chosen by the user. The relevance may be defined in terms of discrimination or coherence. A feature is relevant for discrimination if it is important to separate the selected observations $\mathcal{X}'_{\mathcal{F}'}$ from the unselected observations $\mathcal{X}_{\mathcal{F}'} \setminus \mathcal{X}'_{\mathcal{F}'}$. A feature is relevant for coherence if it contributes to the *compactness* of the set of selected observations $\mathcal{X}_{\mathcal{F}'}$. The relevance metrics are further detailed in Sec. 5. The feature scoring view also allows the user to select a subset of $\mathcal{F}'$ through a rectangular selection widget.

Consider a projection view, created from a feature subset $\mathcal{F}'$, with points colored by the classification ground-truth. Well separated and uniformly colored clusters in this view would be strong evidence that a simple distance-based classifier would be effective. Our visual analytics workflow and associated tool support the guided search for such feature subsets $\mathcal{F}'$, as demonstrated by the example that follows.

4. Application: selecting features for skin lesion classification

The analysis of pigmented skin lesion images by computers is an active research area with almost 30 years of history. One of the most researched problems in this field is automatic melanoma diagnosis [KG12]. Melanoma is a malignant skin cancer that affects the melanocytes, cells responsible for distributing the pigment melanin to other skin cells. Its diagnosis is sometimes difficult, because melanoma can be visually mistaken for commonly occurring benign skin lesions. Clinicians follow well defined criteria to diagnose melanoma, and automatic methods commonly use features that correspond to these criteria [KG12].

Feature selection is often used to develop skin lesion classification systems [KG12], for the reasons already mentioned in Sec. 1. This section describes the use of our visual analytics tool in this context. We consider a subset of the EDRA atlas dataset [GA02] containing 753 dermoscopic color images of manually segmented skin lesions. From these lesions, we extracted $m = 346$ image features, using classical color and texture descriptors found in the literature [KG12]. We grouped image labels, assigned by medical experts, into two classes: melanoma (485 images) and naevi (blue, Clark, combined, congenital and dermal; 268 images).

Assuming the role of a classification system designer, we want to answer the following questions about the data:

1. Which (small) subset $\mathcal{F}' \subseteq \mathcal{F}$, if any, is sufficient to train a classifier as good as one trained using all the features $\mathcal{F}$?
2. How do features compare in discriminative power?
3. Which kinds of images are hard to classify correctly?

To answer these questions, we executed the following workflow using our tool. Firstly, we load the data, project it to 2D by selecting all features ($\mathcal{F}' \leftarrow \mathcal{F}$), and color the points in the plot by their classes (red for naevi, blue for melanoma). We obtain a large overlap between the two classes in the projection plot, as seen in Fig. 1.

Scoring-based selection: Since $\mathcal{F}'$ is large, the best way to reduce it is to employ feature scoring. For this, we select all points of a class (melanoma or naevi) and run a feature ranking technique on the selection $\mathcal{X}'$. Fig. 2a shows the resulting feature scoring view for recursive feature elimination (RFE) [GWBV02]. In this view, each bar in the plot corresponds to a feature. The height of a bar is proportional to its score (higher scores are better), and the color of a bar encodes the distribution of the selected observations inside the range of the feature (yellow represents higher density). We chose to select the 150 highest ranked features as the new subset $\mathcal{F}'$. The resulting projection (Fig. 2b) does not show a better separation between points in our two classes. However, we have already selected less than half of the original features. We proceed by using another feature scoring metric based on an ensemble of randomized decision trees [GEW06]. The scoring view shows significant differences in the relevance of features (Fig. 2c). At this step, we select only the 30 highest scoring features. We start seeing a slight separation between classes in the resulting plot (Fig. 2d). After a few more iterations of

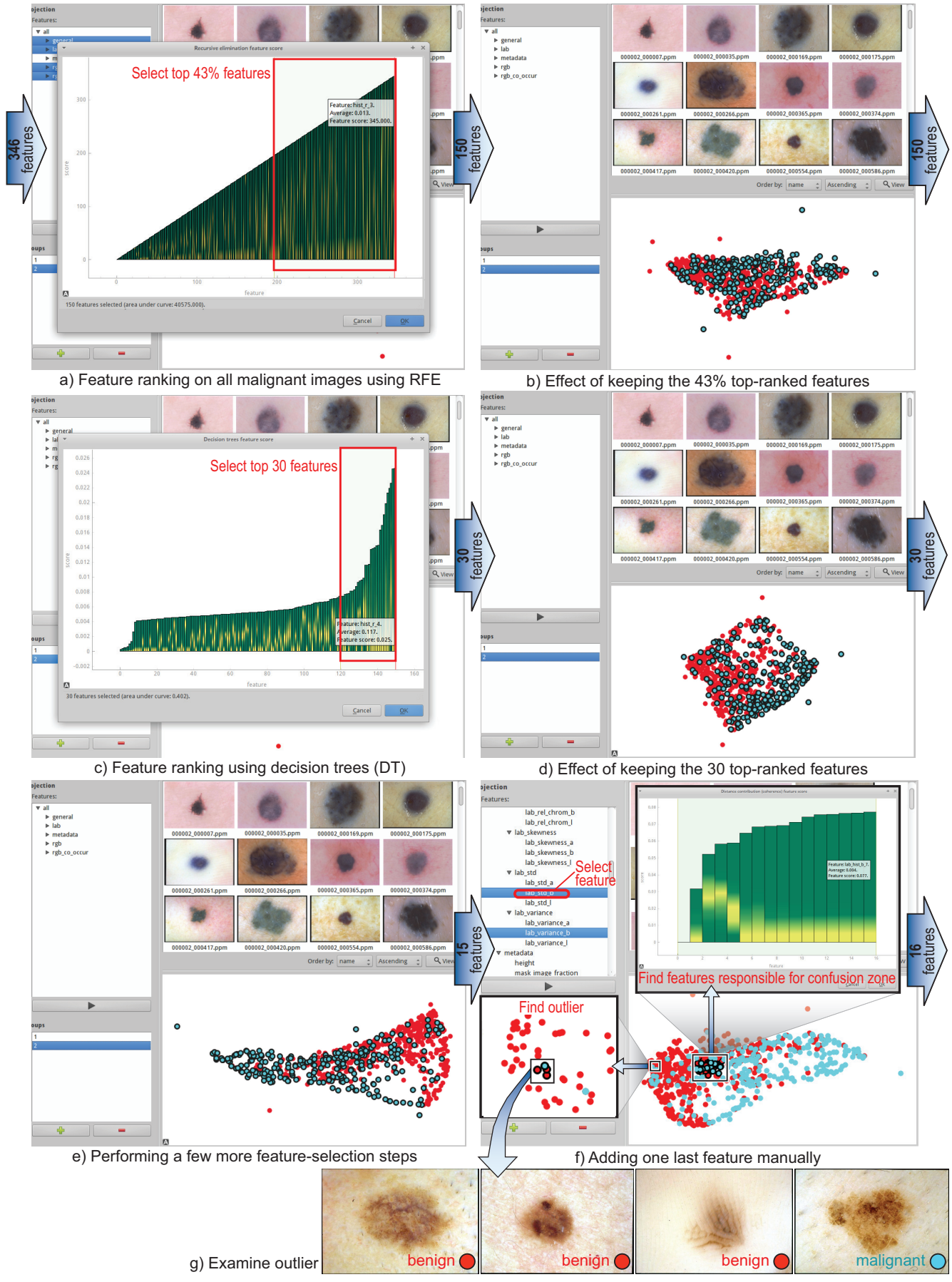

Figure 2: *Visual analytics workflow for reducing feature selection problem in the construction of a skin image classifier (see Sec. 4).*

using different metrics to score and select features, and backtracking whenever necessary, we maintain only 15 features, obtaining the separation between classes shown in Fig. 2e.

Manual refinement: At this point, we became convinced that further iterations of feature scoring and selection did not easily yield better perceived separation between classes in the projection view. However, we can still refine $\mathcal{F}'$. We manually inspect the impact of removing each feature in $\mathcal{F}'$ individually, and also the impact of adding semantically related features, e.g., siblings of the selected features in the hierarchy, and judge the resulting perceived separation. Using this procedure, we obtained the result shown in Fig. 2f, by adding a single sibling feature. At this step, we retained 16 of the original 346 features.

Understanding limitations: Finally, we try to understand what causes the *overlap* between points belonging to different classes. For instance, consider the points selected in Fig. 2f. It is natural to ask which features cause this confusion. We employ coherence-based scoring on this selection, obtaining the results shown in the feature scoring view in Fig. 2f. According to our metric, 11 out of the 16 features contribute almost equally to the compactness of the selected points. Unfortunately, this indicates that it may be hard to eliminate features based on this heuristic.

It is also possible to inspect *outliers*, points whose neighborhood belongs mostly to a different class. We look for outliers in the projection view, and select each of them together with a number of neighbors. The inset in Fig. 2f magnifies such a selection. We use the observation view to inspect the four corresponding images (Fig. 2g): three naevi (red) and one melanoma (cyan). In this case, the images are visually very similar. This is an indication that our current features are not powerful enough to capture the differences between these images. In general, this feedback may lead to the creation of features specifically designed for the problems observed in the data.

Evaluation: To evaluate the effectiveness of our feature selection, we considered four different classification methods: k-nearest neighbors (KNN), random forests (RFC), linear support vector machines (SVML) and radial basis function support vector machines (SVMR). These methods are commonly applied in similar tasks [KG12, ZIL12, CS13]. We performed 5-fold cross validation on our input data while performing grid search on a subset of the parameter space of each method. Table 1 shows the highest average accuracy over the five folds for a given classifier and parameter pair. The table compares the accuracy obtained using all features $\mathcal{F}$ against using the features $\mathcal{F}'$ selected by our exploration. As expected, $\mathcal{F}'$ yields very similar performance, despite retaining less than 5% of the original features in $\mathcal{F}$. We conclude that our feature subset selection was successful and that, if higher scores are desired, new features need to be considered.

	RFC	KNN	SVMR	SVML
All features $\mathcal{F}$	**0.82**	0.78	0.8	**0.79**
Selected features $\mathcal{F}'$	0.81	**0.79**	**0.8**	0.79

Table 1: *Highest average accuracy over all folds for a given classifier and parameter pair (rounded to two decimal places)*

5. Implementation

We implemented our tool in Python, using numpy [vdWCV11], scipy [JO*14], pyqt, matplotlib [Hun07], skimage [vdWSN*14], sklearn [PVG*11], pyqtgraph and mlpy [AVM*12]. For dimensionality reduction, we employed Least Square Projection (LSP) [PNML08], a fast non-linear projection featuring very good distance preservation properties. We employed three classes of feature scoring metrics: univariate (χ^2, one-way ANOVA, and distance-based compactness, which we will describe in future publications), multivariate (IRelief, [Sun07]), and wrappers (ensembles of randomized decision trees [GEW06], randomized linear regression [MB10], and recursive feature elimination (RFE) [GWBV02]). Wrappers tend to be the most reliable feature scoring methods, since they are based on the results obtained by classifiers. On a 3.2GHz Linux PC with 8 GB RAM, our implementation ran all exploration scenarios described in this text in interactive time.

6. Discussion

Several aspects of our proposal are worth emphasizing.

Importance: As already mentioned, selecting a small set of features is beneficial for several reasons. Firstly, computing features over large image collections may be an expensive process. Therefore, reducing the number of features makes the classification pipeline faster. Furthermore, in supervised learning, two well-known classes of errors are bias and variance. Minimizing the number of features is one way of minimizing the variance in the data, which often causes a classifier to be overly affected by small variations in the input. More importantly, the resulting feature set may guide the task of feature design, since the reduced set of feature potentially reveals which features are relevant for classification. This information is highly valuable for the classification system designer, who needs to discover how to improve the system. Our proposal provides visual feedback that is unparalleled by existing feature selection processes.

Generality: Our method is not limited to features extracted from images, and can also deal with categorical features.

Simplicity: Our approach is arguably simple to explain for users with basic machine learning knowledge. The main interactions are also simple: sorting, selecting, loading and saving states. Except for some of the feature scoring metrics, our components are computationally scalable and allow exploration in interactive time.

Limitations: The feedback given by the projection plot is highly dependent on the data and on the quality of the underlying projection technique. We employed non-linear techniques that can preserve distances well in many cases [PNML08, JPC*11, PEB*11], giving an accurate 2D view of mD similarities. We claim that if a 2D projection of the observations restricted to a set of features shows a good separation between points belonging to different classes, a classifier based on distances is likely to succeed. We do not claim the converse, since even if the groups are not separated in the plot, they still might be discriminated by a supervised learning technique. Our feature scoring methods cannot guarantee the selection of an *optimal* feature subset for classification efficacy, since finding such a subset is computationally intractable, and the scoring methods are also not guaranteed to produce visual separations in the plot. However, we allow a simple workflow that enables useful insights for creating effective image classification systems and selecting feature subsets.

7. Conclusions

We proposed a visual analytics approach and tool to help classification system designers in selecting small and effective feature subsets. The selection is guided by dimensionality reduction plots and feature scoring metrics. The approach was evaluated for feature selection in the task of skin lesion image classification. In this use case, our results show that our method allows selecting a very small feature subset that yields the same classification accuracy as a much larger feature set. Our method is generic with respect to data, computationally and visually scalable, and arguably easy to learn and use.

In future work, we intend to integrate information obtained from classifiers into our tool, along with other data summaries to guide feature selection. We are also interested in applying our tool to more general feature selection problems. Finally, we plan on facilitating the exploration of the space of alternatives by employing workflow visualization [SGL09].

We would like to thank CAPES, FAPESP (2012/24121-9, 2011/18838-5) and CNPq (302970/2014-2, 479070/2013-0) for the financial support.

References

[AVM*12] ALBANESE D., VISINTAINER R., MERLER S., RICCADONNA S., JURMAN G., FURLANELLO C.: mlpy: Machine learning in Python, 2012. arXiv:1202.6548. 4

[Ben09] BENGIO Y.: Learning deep architectures for AI. *Found. Trends Mach. Learn. 2*, 1 (Jan. 2009), 1–127. 1

[BTK11] BERTINI E., TATU A., KEIM D.: Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE TVCG 17*, 12 (Dec 2011), 2203–2212. 1

[CS13] CRIMINISI A., SHOTTON J.: *Decision Forests for Computer Vision and Medical Image Analysis*. Springer, 2013. 1, 4

[DB05] DREISEITL S., BINDER M.: Do physicians value decision support? a look at the effect of decision support systems on physician opinion. *Artif. Intell. Med. 33*, 1 (Jan. 2005), 25–30. 1

[GA02] G. ARGENZIANO H.P. SOYER V. D. G. E. A.: *Interactive atlas of dermoscopy*. EDRA Medical Publishing and New Media, Milan, Italy, 2002. 2

[GE03] GUYON I., ELISSEEFF A.: An introduction to variable and feature selection. *J. Mach. Learn. Res. 3* (Mar. 2003), 1157–1182. 1

[GEW06] GEURTS P., ERNST D., WEHENKEL L.: Extremely randomized trees. *Mach Learn 63*, 1 (2006), 3–42. 2, 4

[GWBV02] GUYON I., WESTON J., BARNHILL S., VAPNIK V.: Gene selection for cancer classification using support vector machines. *Machine Learning 46*, 1-3 (2002), 389–422. 2, 4

[Hun07] HUNTER J. D.: Matplotlib: A 2D graphics environment. *Comput Sci Eng 9*, 3 (2007), 90–95. 4

[JO*14] JONES E., OLIPHANT T., ET AL.: SciPy: Open source scientific tools for Python, 2014. URL: http://www.scipy.org/. 4

[JPC*11] JOIA P., PAULOVICH F., COIMBRA D., CUMINATO J., NONATO L.: Local affine multidimensional projection. *IEEE TVCG 17*, 12 (2011), 2563–2571. 4

[KG12] KOROTKOV K., GARCIA R.: Methodological review: Computerized analysis of pigmented skin lesions: A review. *Artif. Intell. Med. 56*, 2 (Oct. 2012), 69–90. 1, 2, 4

[KPB14] KRAUSE J., PERER A., BERTINI E.: INFUSE: interactive feature selection for predictive modeling of high dimensional data. *IEEE TVCG 20*, 12 (2014), 1614–1623. 1

[KSH12] KRIZHEVSKY A., SUTSKEVER I., HINTON G. E.: Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, Pereira F., (Ed.). Curran, Inc., 2012, pp. 1097–1105. 1

[LWBP14] LIU S., WANG B., BREMER P.-T., PASCUCCI V.: Distortion-guided structure-driven interactive exploration of high-dimensional data. *CGF 33*, 3 (2014), 101–110. 1, 2

[LY05] LIU H., YU L.: Toward integrating feature selection algorithms for classification and clustering. *IEEE TKDE 17*, 4 (2005), 491–502. 1

[MB10] MEINSHAUSEN N., BÃIJHLMANN P.: Stability selection. *J. Royal Stat Soc: Series B 72*, 4 (2010), 417–473. 4

[PEB*11] PAULOVICH F., ELER D., BOTHA C., MINGHIM R., NONATO L.: Piecewise laplacian-based projection for interactive data exploration and organization. *CGF 30*, 3 (2011), 1091–1100. 4

[PNML08] PAULOVICH F. V., NONATO L. G., MINGHIM R., LEVKOWITZ H.: Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE TVCG 14*, 3 (2008), 564–575. 4

[PVG*11] PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M., DUCHESNAY E.: Scikit-learn: Machine learning in Python. *J Mach Learn Res 12* (2011), 2825–2830. 4

[SGL09] SHRINIVASAN Y., GOTZ D., LU J.: Connecting the dots in visual analytics. In *Proc. IEEE VAST* (2009), pp. 123–130. 4

[Sun07] SUN Y.: Iterative relief for feature weighting: Algorithms, theories, and applications. *IEEE TPAMI 29*, 6 (2007), 1035–1051. 4

[TFH11] TURKAY C., FILZMOSER P., HAUSER H.: Brushing dimensions: A dual visual analysis model for high-dimensional data. *IEEE TVCG 17*, 12 (2011), 2591–2599. 1

[TMF*12] TATU A., MAAS F., FARBER I., BERTINI E., SCHRECK T., SEIDL T., KEIM D.: Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In *Proc. IEEE VAST* (2012), pp. 63–72. 1

[Tsa12] TSAI C.-F.: Bag-of-words representation in image annotation: A review. *Intl. Scholarly Res. Notices 2012* (2012). 1

[vdWCV11] VAN DER WALT S., COLBERT S., VAROQUAUX G.: The numpy array: A structure for efficient numerical computation. *Comput Sci Eng 13*, 2 (2011), 22–30. 4

[vdWSN*14] VAN DER WALT S., SCHÖNBERGER J. L., NUNEZ-IGLESIAS J., BOULOGNE F., WARNER J. D., YAGER N., GOUILLART E., YU T., THE SCIKIT-IMAGE CONTRIBUTORS: scikit-image: image processing in Python. *PeerJ 2* (6 2014), e453. 4

[YRWG13] YUAN X., REN D., WANG Z., GUO C.: Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. *IEEE TVCG 19*, 12 (Dec 2013), 2625–2633. 2

[ZIL12] ZHANG D., ISLAM M., LU G.: A review on automatic image annotation techniques. *Pattern Recogn 45*, 1 (2012), 346–362. 1, 4