Modeling Problem Transformations based on Data Complexity

Ester BERNADÓ-MANSILLA and Núria MACIÀ-ANTOLÍNEZ
Grup de Recerca en Sistemes Intel.ligents
Enginyeria i Arquitectura La Salle, Universitat Ramon Llull
Quatre Camins 2, 08022 Barcelona (Spain)
{esterb,nmacia}@salle.url.edu
http://www.salle.url.edu/GRSI

Abstract. This paper presents a methodology to transform a problem to make it suitable for classification methods, while reducing its complexity so that the classification models extracted are more accurate. The problem is represented by a dataset, where each instance consists of a variable number of descriptors and a class label. We study dataset transformations in order to describe each instance by a single descriptor with its corresponding features and a class label. To analyze the suitability of each transformation, we rely on measures that approximate the geometrical complexity of the dataset. We search for the best transformation minimizing the geometrical complexity. By using complexity measures, we are able to estimate the intrinsic complexity of the dataset without being tied to any particular classifier.

Keywords. Data Complexity, Classification, Breast Cancer Diagnosis

1. Introduction

The data complexity analysis [1] concerns the study of to what degree patterns can be extracted from datasets. Most of the research in pattern recognition, machine learning, and other related areas have focused on designing competitive classifiers in terms of generalization ability, explanatory capabilities, and computational time. However, limitations in classification performance are often due to the difficulties of the dataset itself. The data complexity analysis is a recent area of research that tries to characterize the intrinsic complexity of a dataset and find relationships with classifier's accuracy.

This paper uses the data complexity analysis to study problem transformations for a particular case of breast cancer diagnosis. The original problem is not directly tractable by a classifier, since each patient has a variable number of descriptors. Thus, prior to the application of any classifier, the descriptors must be synthetized into a single one. Several synthetic representations are possible, each leading different classification results. Moreover, due to the nature of the problem, the domain experts are unable to specify the best one. A previous approach was to use the classifier's accuracy as a measure of quality of the different synthetic representations. A limitation of the approach was that the classifier's accuracy depended both on classifier's bias and dataset characteristics, misleading the selection of the best method and confusing the interpretation provided to the human experts. We consider the characterization of data complexity as a method to select the best synthetic representation. The proposed methodology sets a framework that guides us in the selection of problem transformations without being tied to any particular classifier. Our results are also meaningful to the domain experts, because they provide intrinsic information of the problem at hand.

The paper is structured as follows. Section 2 reviews the data complexity analysis and its application to classification problems. Section 3 describes the problem we address in more detail. Next, we study problem transformations and their charaterization by means of the complexity analysis. We present the results and finally, we summarize the conclusions and future work.

2. Analysis of Data Complexity

The complexity of a classification problem can be attributed to three main sources [2]. *Class ambiguity* is identified as the difficulty given by non-distinguishable classes. This may be due to the intrinsic ambiguity of the problem, or to the fact that the features are not sufficient to discriminate between the classes. Class ambiguity sets a lower bound on the achievable error rate, which is called Bayes error. The *sparsity of the training set* is related to the number and representativity of the available instances. The *boundary complexity*, the third source of difficulty, can be characterized by the Kolmogorov complexity, or the minimum length of a computer program needed to reproduce the class boundary. Class ambiguity and training set sparsity are properties of the specific dataset. Once the dataset is fixed, these complexities are irrecoverable. On the other hand, the geometrical complexity is more relevant to the study of classifier's behavior. The characterization of the boundary complexity may be useful to explain the different performance of several classifiers on a given dataset. That is why recent studies on data complexity have been mainly focused on the boundary complexity.

2.1. Applications of Data Complexity Analysis

One of the primary goals of the data complexity analysis was to understand the classifier's performance. The idea arose from the difficulty of traditional studies to understand the differences among several classifiers when compared on several datasets. The common table with comparison of accuracies did not reveal the reasons why a given classifier performed better or worse than others for certain problems. Thus, some studies tried to characterize the dataset complexity and relate it to the classifier's performance. By doing so, one could build a model of classifier's accuracy based on dataset complexity and use the model to set expectations of classifier's accuracy. Some studies in this direction found linear correlations between the classifier's error and some measures of complexity [3,4]. Other investigations attempted to find domains of competence of several classifiers in a space defined by complexity measures [5,6]. Thus, given a new problem characterized by its complexity, the model could be used as a guide for classifier selection. Often, a learning algorithm has different configurations available. Therefore, the same methodology could be used to extract rules for classifier's adaption to a particular problem.

The data complexity analysis can also be used at the preprocessing stages of classification, such as in prototype selection [11,10] and feature selection. There, the characterization of the dataset is employed to select a suitable problem with reduced dimensionality. This paper explores the application of the data complexity analysis to find proper transformations of a classification problem.

2.2. Data Complexity Measures

Ho & Basu [2] proposed a set of metrics estimating data complexity. The metrics are classified in four categories as follows.

2.2.1. Overlap of Individual Feature Values

These metrics evaluate the power of individual attributes to discriminate between classes.

Maximum Fisher's discriminant ratio (F1): For each attribute, the Fisher's discriminant ratio is calculated as: $f = (\mu_1 - \mu_2)^2/(\sigma_1^2 + \sigma_2^2)$, where μ_1, μ_2 and σ_1^2, σ_2^2 are the means and variances of the attribute for each of the two classes, respectively. The metric uses the most discriminant feature as the one having the maximum Fisher's value.

Volume of overlap region (F2): The overlap region of a feature is computed as the overlap range divided by the total range of that feature. F2 is the product of the overlap regions of each attribute.

Feature efficiency (F3): It describes to what extent each feature contributes to the class separation. It consists in removing the ambiguous instances (i.e., those instances belonging to different classes that fall in the overlapping region) for each feature. The efficiency of each feature is the ratio of the remaining non-overlapping points to the total number of points. The largest feature efficiency of all features is taken as F3.

2.2.2. Separability of classes

This family of metrics takes into account the dispersion of classes in the feature space based on neighborhood distances.

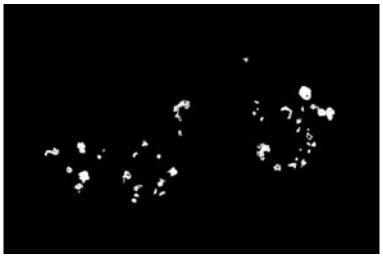
Length of class boundary (N1): It refers to the percentage of points in the dataset that lie in the class boundary [9]. Firstly, we generate the minimum spanning tree (MST) connecting all training samples, using the Euclidian distance between each pair of points. Then, we compute the fraction of points joining opposite classes to the total number of points. This measure is sensitive to the separability of classes and the clustering tendency of points belonging to the same class.

Ratio of average intra/inter-class nearest neighbor distances (N2): For each point, we calculate its nearest neighbor point belonging to the same class and the nearest neighbor belonging to the opposite class. Then, the averaged distances connecting intra-class nearest neighbor points are divided by the averaged distances of inter-class nearest neighbors.

2.2.3. Geometry of Class Manifolds

It evaluates the overlap between classes and how the classes are distributed as hyperspheres in the feature space. This is more related to the interior descriptions of geometry.

Nonlinearity (N4): It estimates a convex hull for each class by linear interpolation of randomly drawn pairs of points from the same class. Then, a nearest neighbor classifier is trained with the original training set and tested with the extended set of points approximating the convex hull. N4 is the error of the classifier.



Area	Number of holes
Perimeter	Convex perimeter
Compactness	Roughness
Box Min. X,Y; Max. X,Y	Length
Feret (min. bounding box) X,Y	Breadth
Feret minimum diameter	Elongation
Feret maximum diameter	Centroid X,Y
Feret mean diameter	Angle of principal axis
Feret elongation	Angle of secondary axis

Figure 1. Example of a mammographic image with several microcalcifications (left) and set of features extracted from each microcalcification (right)

2.2.4. Sparsity

The sparsity is estimated as the number of points to the number of dimensions.

3. The Problem

The problem addressed in this paper consists in the breast cancer diagnosis based on the features extracted from mammographic images. Mammographic images may contain microcalcifications, which are tiny specks of mineral deposits (calcium), that can be found scattered or clustered throughout the mammary gland. The specks may either indicate the presence of tiny benign cysts or early breast cancer. In the latter case, studies reveal that shapes and sizes of microcalcifications are relevant features to determine if they constitute high risk of malignant cancer.

The dataset we used was obtained from mammograms collected by *Dr. Josep Trueta* University Hospital, whose diagnosis was known from biopsies. Each mammogram was digitized and later processed [7]. The result was a set of 216 instances, where each instance belonged to a mammogram with variable number of microcalcifications and a class label that corresponded to the diagnosis. Each microcalcification was characterized by 23 features mostly describing shapes and sizes. Figure 1 shows an example of a mammogramphic image and the list of features obtained for each microcalcification. For more details, see [7].

In this application, an instance-based learner was used [12]. The reason was that the most similar images to the new case were presented to the human expert as a way of explaining the diagnosis. Due to the variable number of microcalcifications present in each instance, the problem was untractable directly by the classifier. Early works averaged the features of all microcalcifications as a synthetic case [12]. Although the classification performance reached the same accuracy as the human experts, there was uncertainty about the correct method of synthetizing the different microcalcifications. We wondered whether other methods could be used that were easier for the human experts when performing a visual inspection of the mamographic image.

A possible procedure is to test the error of the classifier with the datasets obtained from different types of transformations and then, choose the best transformation as the one with the minimum error. However, this can lead to conclusions too tied to the classifiers applied. That is, the error of the classifier is influenced both by the complexity of the dataset and the proper design of the classifier. By studying the complexity of the dataset, we can provide a more theoretical framework. Our proposal is to characterize the

Table 1. Problem transformations

Method	Synthetic case
Average	Average of all attributes
Centroid	Feature values belonging to the centroid of the cluster
Random	A random microcalcification
All	Each microcalcification constitutes a different case
Min. area	The microcalcification with the minimum area
Max. roughness	The microcalcification with the maximum roughness
Max. compactness	The microcalcification with the maximum compactness
Max. elongation	The microcalcification with the maximum elongation
Max. feret elongation	The microcalcification with the maximum feret elongation
Max. holes	The microcalcification with the maximum number of holes

complexity of each dataset resulting from the different transformations and then select the best transformation based on the minimum complexity.

4. Results

Table 1 describes the different problem transformations we analyzed to synthetize a number of microcalcifications into a single one. The average approach involves all microcalcifications present in the mammogram and computes their average feature values. These are the feature values used as the synthetic case. The centroid approach computes the centroid point of all microcalcifications, considering Euclidean distances. We included a random selection of a microcalcification to test whether the selection of a particular microcalcification was relevant. We also included the case where all microcalcifications were used as different cases, each labeled with the class corresponding to the global diagnosis given by the biopsy. The rest of methods select a particular microcalcification, based on the value of a single feature. We based our selection on an early analysis of data [8] that revealed that the most relevant set of features for cancer discrimination were: the area, compactness, feret elongation, number of holes, roughness, and elongation. Thus, the different transformations selected a single microcalcification which were: the microcalcification with the minimum area, the maximum roughness, the maximum compactness, the maximum elongation, the maximum feret elongation, and the maximum number of holes, respectively.

Table 2 shows the values of the complexity metrics computed for each problem transformation. See that each transformation is characterized by a complexity space of 6 metrics, where F1, F2, and F3 evaluate the discriminative power of features, N1 and N2 consider the separability of classes, and N4 the nonlinearity of class boundary. The rows are sorted in ascending order of metric N1.

To our understanding, the metrics that compute the discrimination of individual features are not very relevant to the complexity estimation. For example, a high value of F1 indicates that an attribute discriminates well and consequently, the problem should be easy. However, a small value for F1 does not necessarily imply a difficult problem. We should look at the rest of the metrics to complete our estimation. The values obtained in F1, F2, and F3 are very close in all transformations. The exception is the *Average* approach, which obtains simultaneously a high value in F1 and F3 (see figure 2(a)). This

Table 2. Complexity of problem transformations (columns F1 to N4) and error of nearest neighbor classifiers (columns 1-NN and 3-NN). The most complex problem according to the given metric is marked in bold, and the easiest problem is marked in itallic.

Dataset	F1	F2	F3	N1	N2	N4	1-NN	3-NN
Average	*0.3131*	0.0081	*0.1019*	*0.528*	*0.899*	0.9653	0.370	0.347
Max. holes	0.0976	0.0039	0.0463	0.574	0.927	0.9730	0.398	0.394
Random	0.0356	*0.0015*	0.0509	0.588	0.961	0.9614	0.421	0.394
All	**0.0150**	0.0080	**0.0052**	0.593	0.928	*0.9516*	0.411	0.394
Max. roughness	0.0878	0.0044	0.0463	0.616	0.955	0.9730	0.421	0.458
Centroid	0.0700	0.0072	0.0463	0.620	0.972	0.9653	0.458	0.458
Max. elongation	0.0969	0.0027	0.0509	0.630	0.967	0.9691	0.472	0.449
Max. compact.	0.0969	0.0027	0.0509	0.630	0.967	0.9691	0.472	0.449
Max. feret elong.	0.2518	0.0108	0.0463	0.676	**1**	0.9769	0.468	0.421
Min. area	0.1379	**0.0172**	0.0556	**0.704**	0.979	**0.9846**	0.472	0.454

could indicate that its complexity is low. The *All* transformation gives the smallest values in F1 and F3. On the other hand, the two largest values of F2 (since F2 is the volume overlap, a large value of F2 could mean a difficult problem) correspond to *Max. feret elong.* and *Min. area*. Both transformations also give the largest values in N1 (see figure 2(b)). This couple effect may point that these are the most difficult datasets.

Metrics focused on the distribution of classes should convey more information on complexity. That is why we sorted the rows by N1. Column N1 gives the values of the length of class boundary and N2 the fraction of intra/inter-class distances. Regarding the length of class boundary, the *Average* appears as the best transformation. This also agrees with the values obtained by *Average* in F1 and F3. The worst approximations are those that select a single microcalcification based on individual feature values (except for the number of holes). Surprisingly, the approaches of using all microcalcifications and selecting a random microcalcification give boundary lengths rather small (i.e., low complexity). This means that all microcalcifications carry similar information for carcinoma detection, which justifies why the average approach appears as the best transformation. N1 and N2 are fairly correlated, as shown in plot 2(c). This is reasonable because the two metrics test the class dispersion by using nearest neighbor distances.

Finally, N4 checks for nonlinearity. Under N4, the worst transformation is *Min. area*, which agrees with the result of N1, N2, and F2.

In general, the metrics agree that the worst transformations are *Max. feret elong.* and *Min. area*, while the *Average* appears as the best tranformation.

To validate whether our complexity estimation corresponded to the complexity as seen by the classifiers, we tested two k-nearest neighbors (k-NN) classifiers and computed their errors for each transformation. The classifiers' error was estimated by an stratified 10-fold cross-validation procedure. The two last columns of table 2 show the error of the nearest neighbors (k=1 and k=3 respectively) for each transformation. See that the error of the classifiers tends to rise for increasing values of N1. Also note that the worst complex problems (as predicted by the complexity metrics) correspond to the highest classifiers' error, while the easiest problems correspond to the smallest errors. Figures 2(e) and 2(f) show graphically the correlation between the classifiers' error and metric N1.

In [4] a linear correlation was observed between some classifiers' error and metrics N1 and N2. If we applied the model derived in the paper, for N1 ranging from 0.528 to

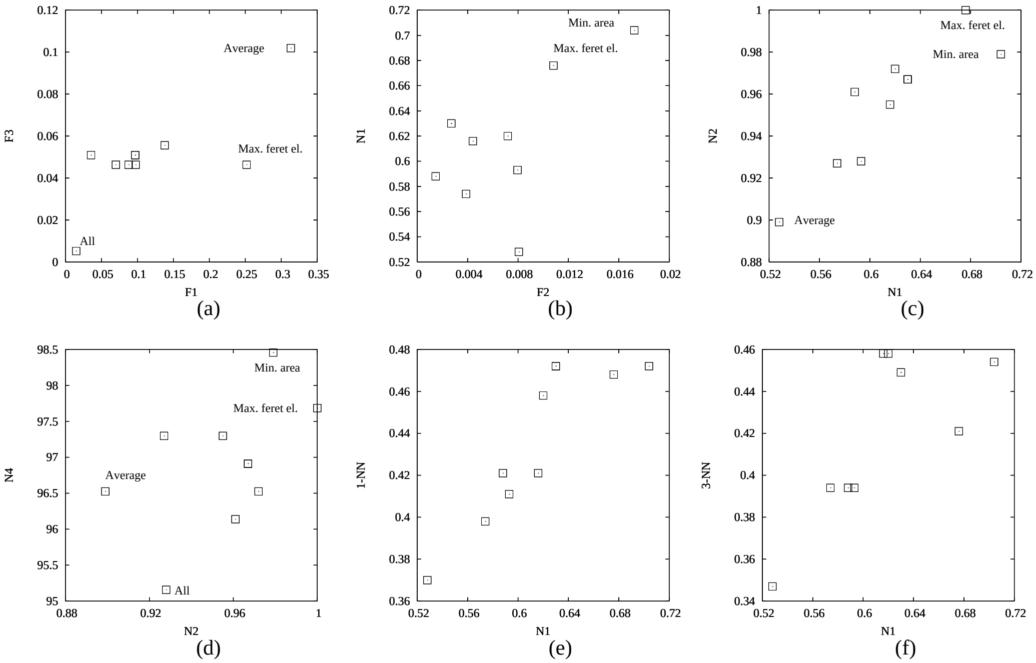


Figure 2. Complexity projections and error of classifiers. Plots (a)-(d) show each transformation plotted in several two-dimensional projections of the complexity measurement space. Plots (e)-(f) show the error of 1-NN and 3-NN vs. metric N1.

0.704, we would predict a classifier's error in the interval [0.3,0.5]. See that the results obtained by our classifiers fit correctly this model, which confirms the suitability of the complexity measurement space to set estimations of the classifier's error.

5. Conclusions

The data complexity analysis has provided a theoretical framework to select and justify the best transformation for the problem of breast cancer diagnosis based on features extracted from a variable number of microcalcifications present in a mammogram image. By selecting the transformation that minimizes data complexity, we are able to increase the generalization ability of classifiers and thus, give better support to the breast cancer diagnosis. Besides, the analysis may be meaningful to experts, since it provides guidelines on how to inspect visually the microcalcifications.

The estimation of complexity in all the studied transformations has been fairly high, which consequently has lead to a high classifier's error. The results reveal that the breast cancer diagnosis relying only on microcalcifications reach a maximum accuracy rate that seems difficult to be overcome, regardless of the efforts for transforming the dataset or using better adapted classifiers. Probably some attributes are missing, specifically those relating patient's age, antecedents, etc. (as considered by the experts). It would be interesting to see how the data complexity decreases with the addition of such new features.

Our results may be also limited due to a sparse sample. In fact, the dataset contained only 216 instances. The finite and sparse samples limit our knowledge about the geometrical complexity, thus we are addressing only the apparent complexity of a problem based on a given training dataset.

The study of data complexity provides expectations on the error of classifiers. Thus, besides searching for the best classifier solving a particular problem, we may seek for problem transformations that present data in a more learnable way. A similar methodology could be adapted to processes such as feature extraction, selection, and aggregation. The data complexity analysis has also served to identify the domains of competence of classifiers in the complexity measurement space. As a further step, one could use these studies to transform a given dataset for a particular type of classifier; i.e., translating the problem to the domain of competence of the classifier of interest.

Acknowledgments

The authors thank the support of *Enginyeria i Arquitectura La Salle*, Ramon Llull University, as well as the support of *Ministerio de Ciencia y Tecnología* under projet TIN2005-08386-C05-04. The authors acknowledge also J. Martí for providing the dataset, in the context of projects FIS 00/0033-02 and TIC2002-04160-C02-02.

References

[1] Basu,M., Ho,T.K.: Data Complexity in Pattern Recognition, Springer (2006)

[2] Ho,T.K. and Basu,M.: Complexity Measures of Supervised Classification Problems, IEEE Trans. on Pattern Analysis and Machine Intelligence, **24**:3 (2002) 289-300

[3] Ho,T.K.: A Data Complexity Analysis of Comparative Adavantages of Decision Forest Constructors, Pattern Analysis and Applications, **5** (2002), 102-112

[4] Bernadó-Mansilla,E., Ho,T.K.: Domain of Competence of XCS Classifier System in Complexity Measurement Space, IEEE Trans. on Evolutionary Computation, **9**:1 (2005) 82-104

[5] Ho,T.K., Bernadó-Mansilla,E.: Classifier Domains of Competence in Data Complexity Space, In: Basu, M., Ho,T.K.(eds.): Data Complexity in Pattern Recognition, Springer (2006) 135-154

[6] Bernadó-Mansilla,E., Ho,T.K.: On Classifier Domains of Competence, Proceedings of the 17th International Conference on Pattern Recognition, Vol.1 (2004) 136-139

[7] Martí,J. [et al]: Shape-based feature selection for microcalcification evaluation, Proceedings of the SPIE Medical Imaging Conference on Image Processing Vol.3338 (1998) 1215-1224

[8] Barceló,C., Thió,S.: Estudio piloto sobre el diagnóstico de benignidad o malignidad de las microcalcificaciones mamarias mediante digitalización y análisis estadístico, Technical Report, Secció d'Estadística i Anàlisi de Dades. Departament d'I.M.A. (1997)

[9] Friedman,J.H., Rafsky L.C.: Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests, The Annals of Statistics **7**:4 (1979) 697-717

[10] Mollineda, R.A. and Sánchez, J.S. and Sotoca, J.M., Data characterization for effective prototype selection, Proc. of the 2nd Iberian Conf. on Pattern Recognition and Image Analysis, (2005) 27-34

[11] Singh, S., PRISM - A novel framework for pattern recognition, Pattern Analysis and Applications, **6** (2003) 134-149

[12] Martí, J., Freixenet, J., Raba,D., Bosch,A., Pont,A., Español,J., Bassaganyas,R., Golobardes,E., Canaleta,X., HRIMAC - Una Herramienta de Recuperación de Imágenes Mamográficas por Análisis de Contenido para el Asesoramiento en el Diagnóstico de Cáncer de Mama, Actas del VI Congreso Nacional de Informática de la Salud (2003)