Resolution transfer in cancer classification based on amplification patterns

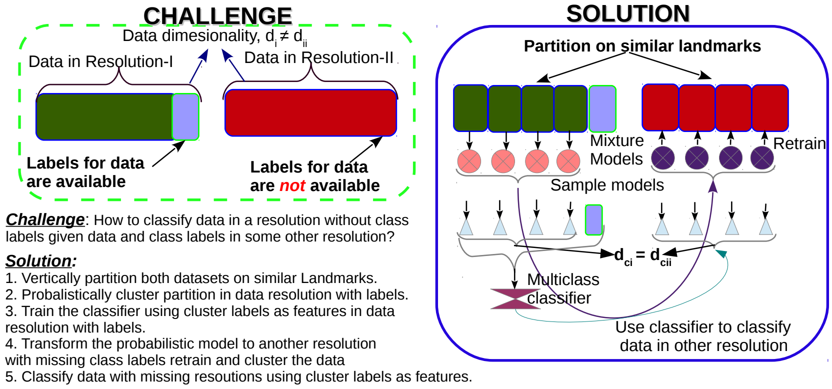
Prem Raj Adhikari[1,2] and Jaakko Hollmén[1]

[1]Helsinki Institute for Information Technology HIIT and
Department of Information and Computer Science,
Aalto University School of Science,
PO Box 15400, FI-00076 Aalto, Espoo, Finland
[2]Department of Physiology and
Turku Center for Disease Modeling
Institute of Biomedicine, University of Turku
Kiinamyllynkatu 10, FI-20520 Turku, Finland
prem.adhikari@utu.fi,jaakko.hollmen@aalto.fi

Abstract. In the current scientific age, the measurement technology has considerably improved and diversified producing data in different representations. Traditional machine learning and data mining algorithms can handle data only in a single representation in their standard form. In this contribution, we address an important challange encountered in data analysis: what to do when the data to be analyzed are represented differently with regards to the resolution? Specifically, in classification, how to train a classifier when class labels are available only in one resolution and missing in the other resolutions? The proposed methodology learns a classifier in one data resolution and transfers it to learn the class labels in a different resolution. Furthermore, the methodology intuitively works as a dimensionality reduction method. The methodology is evaluated on a simulated dataset and finally used to classify cancers in a real–world multiresolution chromosomal aberration dataset producing plausible results.

1 Introduction

Over the years, the measurement technologies have improved considerably providing an opportunity to measure the finer details of the phenomenon [8]. Multiresolution data is generated when the same phenomenon is measured in different levels of detail [13]. The older generation technologies measure only the coarser units of the phenomenon resulting in the data in the coarse resolution while the newer generation technology can measure the finer units of the phenomenon generating the data in the fine resolution. The fine resolution data carries more information in the data sample compared to the coarse resolution data but also has the larger data dimensionality than the coarse resolution data. The importance of combining multiple data sources, and information within a


CHALLENGE
Data dimesionality, $d_i \neq d_{ii}$
Data in Resolution-I
Data in Resolution-II
Labels for data are available
Labels for data are *not* available
SOLUTION
Partition on similar landmarks
Mixture Models
Retrain
Sample models
$d_{ci} = d_{cii}$
Multiclass classifier
Use classifier to classify data in other resolution

Challenge: How to classify data in a resolution without class labels given data and class labels in some other resolution?

Solution:
1. Vertically partition both datasets on similar Landmarks.
2. Probalistically cluster partition in data resolution with labels.
3. Train the classifier using cluster labels as features in data resolution with labels.
4. Transform the probabilistic model to another resolution with missing class labels retrain and cluster the data
5. Classify data with missing resoutions using cluster labels as features.

Fig. 1. The research challenge considered in this contribution and the proposed solution.

single analysis and the availability of multiresolution data in different application areas, such as, image processing, and time–series analysis have given major impetus to the research in multiresolution data analysis [13].

In this contribution, we address an important challange that lies in between supervised learning and learning from unlabeled data; which resembles the semi–supervised learning [5] and the transfer learning [10]. The proposed methodology learns the classifier in one resolution where the class labels are available and transforms the classifier to other resolutions where the class labels are not available. The methodology uses a combination of an unsupervised probabilistic clustering and a supervised multiclass classification in a pipeline to address the challenges in classifying the data in different resolutions when class labels are not available in all the resolutions of the data. In this contribution, we do not propose a new probabilistic modelling algorithm or a multiclass classification algorithm. The novelty in the contribution comes from the design of the pipeline for classifying the datasets in different resolutions by resolution transfer of the classifier and intuitive dimensionality reduction achieved through the unsupervised probabilistic clustering. While the clustering results have been used to improve the classification results [7], such a methodological pipeline of resolution transfer has not been proposed in the literature.

The Figure 1 shows that between the two high dimensional datasets in two different resolutions, only the data in one resolution has the associated class labels; while the class labels are missing in the data in other resolutions. Therefore, the challenge to learn a classifier on the data in the resolution having the class labels and use the same classifier to classify the samples in the data resolution without the associated labels. A simplified approach of using projection methods such as principal component analysis (PCA) would not produce expected results because the representation of data would be lost.

As shown in the bottom panel of Figure 1, first, we vertically partition both the datasets based on some predefined landmarks in such a way that the number of vertical partitions in each dataset in different resolutions are the same. Second, we train the mixture models in each of the vertical partitions of the dataset with the associated labels. Third, we transform the trained mixture models to another resolution with missing class labels. The transformations are performed using the apriori knowledge of the relationships among different resolutions of the data from the domain ontology. Fourth, we retrain the mixture models with the data partitions in data resolution not consisting of associated labels. The retrained mixture model generates the cluster labels for the data partitions in the data resolution with missing class labels. Fifth, we concatenate the obtained cluster labels in both the data resolutions separately to regenerate the whole genome but with a reduced dimensionality. Sixth, we learn a multiclass classifier for the whole genome in the data resolution having the associated class labels. Finally, we can use the same classifier to classify the data resolution not consisting of associated labels. Since the clusters labels which are used as features are equivalent in both the resolution, the classifier can be used for data in both resolution producing comparable results in both resolutions.

2 Methodology of Multiresolution Multiclass Classification

In our proposed methodology, we first set aside the class labels and vertically partition the feature space in both the given high dimensional datasets on specific landmarks in such a way that specific relationship between different data resolutions can be easily established. Furthermore, vertical partition should be such that data in different resolutions will have equal number of partitions while data dimensionality in each partition can be different. We then use unsupervised probabilistic algorithm, i.e., mixture models, to cluster each partition of the data separately. The number of clustering experiments is equal to the number of vertical partitions in the data. We use model selection to determine the number of clusters in each of the partition of the data separately using ten–fold cross–validation as in [12].

The cluster labels generated by the clustering algorithms are then vertically concatenated forming a new dataset with reduced dimensionality for classification. The newly formed dataset obtaied by concatenating the cluster labels emulates the original data but results in the reduced data dimensionality. This is because a cluster label comprise multiple data dimensionality in the vertical partitions of the data thus ameliorating the problem of curse of dimensionality [4].

In our previous research, we have shown that the mixture models learned in a resolution can be transformed to a different resolution provided there exists a well–defined relationship among the different data resolutions [2]. We can use the domain ontology to determine the relationship between the model parameters in different resolutions of data and exploit that to transform the mixture models across different resolutions.

$$\theta_f \sim N(\mu = \theta_c, \sigma = 0.01) = \begin{cases} \theta_f & \text{if } 0 \leq \theta_f \leq 1 \\ \theta_c & \text{if } \theta_f < 0 \text{ or } \theta_f > 1 \end{cases} \tag{1}$$

We can re–sample the number of parameters required in the fine resolution from a normal distribution with the mean (μ) equal to the parameter value in the coarse resolution and a small standard deviation (σ); 0.01 in our experiments. Mathematically, we can represent the transformations as in Equation 1, where θ_c, and θ_f denote the parameters of the mixture components θ in the coarse and the fine resolution. We can also further ensure that the re–sampled parameters obey the laws of probability in such a way that the value of the parameter in the fine resolution is between 0 and 1, i.e., $0 \leq \theta \leq 1$. If the re–sampled value of θ is outside the given range, we replace it with the value of the parameter, θ, in the coarse resolution such that $\theta_f = \theta_c$. Finally, the transformed mixture model is then retrained on the fine resolution data.

We represent the categorical cluster labels as binary features as discussed in [6]. The number of bits in binary features is equal to the number of components in the mixture model for that data partition, i.e., the clusters in the data. For example, if the number of components is four then the clusters one, two, three, and four are represented as: 1000, 0100, 0010, and 0001. We then vertically concatenate the cluster labels in binary representation to represent an entire dataset. This clustering labels can be assumed to be the summary of the patterns present in the data. Finally, a multiclass classifier, e.g., support vector machines, can be trained using on the dataset generated by concatenating the clustering labels.

3 Experiments on Multiresolution Chromosomal Aberrations Dataset

Two chromosomal amplification datasets were available in two different resolutions for our experiments. The data in coarse resolution describing the DNA amplification patterns of 4590 cancer patients were available from [9]. Similarly, data describing the DNA amplification patterns in fine resolution were available from [3]. The coarse resolution data describes the chromosomal amplifications dividing genome in 393 different parts as described in [11]. In contrast, the fine resolution data describes the chromosomal amplifications dividing the genome in 862 different parts. In addition to resolution, another important difference between the datasets is that the coarse resolution data have associated class labels, i.e., the 4590 patients were associated with 73 different cancer types whereas the data in the fine resolution do not have the associated cancer types (class labels).

Data Preprocessing The number of cancer types in the coarse resolution dataset (73) were too high to learn any credible cancer classifier. Some of the cancers had less than 10 samples making it difficult to learn a classifier that generalizes better on the unseen data [5]. Therefore, we only experimented with top 34 cancer types. The top 34 cancer types were chosen because they covers 90% of the data. This simplification reduces the number of samples in the data to 4104. The cancer with the highest number of samples is Neuroepithelial tumors with 544 samples. In contrast, the cancer with the minimum number of samples is Pulmonary sarcoma with only 30 samples. The simplified data is then processed chromosome–wise, i.e., the data describing the genome is vertically partitioned into 24 different chromosomes. When the data is divided into chromosomes, some samples in some chromosome do not show any amplification. We remove those samples without amplifications (vectors with all zeros) because they carry no information about the cancer and also further simplify the experimental procedure.

Chromosome–wise Mixture Modelling After the data have been vertically partitioned into the different chromosomes, we learn the mixture models based on a model selection procedure in a ten–fold cross–validation setting as discussed in [12]. The model selection procedure selects different number of components in the mixture model to fit the data in different chromosomes. Mixture models are generally used to represent the probability distribution of the data. Nevertheless, it can also be used to cluster the data into hard partitions. The number of partitions is equal to the number of components in the mixture model. The cluster labels are then transformed to binary format and chromosomes are concatenated to form the whole genome. We do not use model selection algorithm on Chromosome Y because of lack of data samples. In chromosome Y, the cluster label is 1 if any of the chromosomal regions is amplified; otherwise 0.

Cancer Classification Using SVMs The cluster labels are transformed to binary format and the chromosomes are concatenated to regenerate the whole genome. In the cancer samples showing no amplifications in specific chromosomes that were ignored duriing mixture modeling are replaced by all zeros in the binary features. The example, of four clusters discussed in Section 2, it would be represented as 0000. We then learn multiclass support vector machines using open–source libsvm software package [6]. The kernel type selected is radial basis function. The parameters of the support vector machines are γ and C which also learned in a ten–fold cross–validation setting by a grid search. The support vector machines are initially learned on the original data with full set of features and also on the vertical concatenation of cluster labels. The original data dimensionality is 393 whereas vertical concatenation of cluster labels results in data dimensionality of 112. The concatenation of features result in reduction of data dimensionality that is less than one third of the features in original data dimension. Naturally, when the data dimension is reduced, the accuracy of classification decreases. Figure 2 shows that decrease in accuracy is negligible when the partitions of data is represented by the cluster labels. The computational and memory efficiency of reduced data dimensionality surpasses the decrease in the classification accuracy.

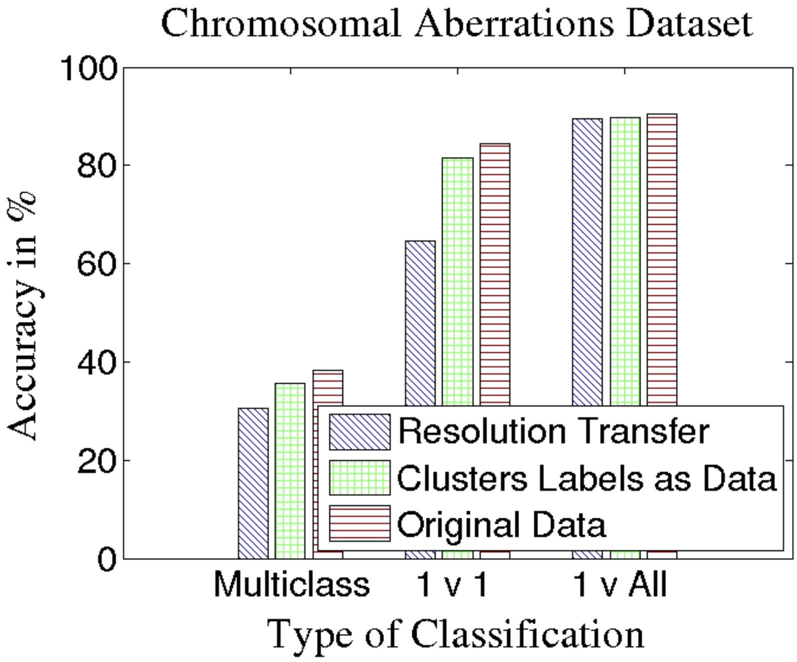


Fig. 2. Comparative study of the accuracy of SVM in different modes of the classification settings.

Resolution transfer of the classifier The crux of this contribution is the resolution transfer of the classifier. We use the knowledge of domain ontology relations to transform the parameters of the mixture model learned in the resolution having the associated class labels to the data resolution not having the associated class labels as in [2]. The transformed model is retrained on the data in other resolution in such a way that the components are not much different than the model in the original resolution. This requirement is enforced because the clustering algorithm should produce same labels for similar data vectors as we are using the cluster labels as data features for classification. Since the algorithm is trained on the data in other resolution, the features in the concerned resolution must be the same as the features in original resolution. Finally, the data in concerned resolution is classified using the classifier trained in original data resolution.

The data in fine resolution obtained from [3] does not have associated cancer types, i.e., class labels. Therefore, we can use the same classifier to classify cancers but we cannot access the performance of the classification algorithm. Therefore, we transform the data to another resolution using deterministic methods similar to the one suggested in [1]. The data in fine resolution can then be classified using the classifier learned in the resolution having associated cancer labels. Furthermore, the performance can be accessed because the transformed data have labels from the coarse resolution.

Figure 2 depicts the classification accuracy of the classifier in different settings. As expected the classification accuracy is best on the original data. The figure shows that the performance of the classifier degrades in resolution transfer. The decrease in performance is expected because the classifier is learned on data in different resolution. The results obtained are promising because resolution transfer provides additional facilities to classify data in different resolution for which the class labels are not available. Furthermore, the negligible decrease in performance can also be attributed to curse of dimensionality [4]. In addition, the performance of multiclass classification is less than 40% which is comparatively less but considering that there are 34 classes, the accuracy is plausible because random classifier would generate accuracy of less than 10%. If all samples are classified as the cancer with the highest number of samples, i.e., Neuroepithelial tumors, the accuracy would be approximately 13%. Therefore, the performance achieved by our methodology is plausible and provides a novel methodology to classify cancers across different resolutions.

3.1 Simulated Data

We also evaluated our methodology on a simulated data set. The simulated dataset was simple with 1000 data samples and 5 dimensions. We randomly sampled a number between 1 and 4, and generate row for the data sample where each element in the sample is equal to the randomly sampled number. For example, if we sample number 3, all the elements in the row are 3. We continue this process until 1000 data samples have been achieved. We then consider first variable as the class and remaining 4 variables as the features. We convert the four variables to binary using decimal to binary conversion system with 3 bits such that 4 dimensional data are transformed into 12 dimensional 0–1 vectors. We then randomly flip the bits of 1200 (10%) data elements to add noise to the dataset. Similarly, we vertically concatenated the 12 dimensional data each dimension one by one to generate 24 dimensional data.

We then group each digit separately again into four groups to run the clustering experiments. Since, we know the number of clusters in the data, i.e., 4, we do not run model selection algorithm in this case. We then evaluate our methodology on this simulated multiresolution data in the similar vein as in Section 3. In this experiment, there was larger discrepancy in classification accuracy in multiclass classification. The original algorithm as well as the clusters labels used as class labels produced accuracy nearing 97% while the resolution transfer produced classification accuracy nearing 75%. Despite the addition of noise, the data is overly simple and in such simple datasets, classifiers often overfit. In one vs one and one vs the rest experiments all the methods produced plausible accuracy of 98.5%.

4 Summary and Conclusions

In this contribution, we were interested in transferring the classifier learning across different resolutions. In our setting, we had access to class labels only in one resolution while the class labels were missing in other resolutions. We learn the classifier in the resolution with the class labels and transfer the learned classifier to classify the data in resolutions with missing class labels. Furthermore, our proposed methodology intrinsically reduces the data dimensionality to less than one–third in the coarse resolution and to less than one–eighth in the fine resolution as an added advantage of the proposed resolution transfer. We experimented our methodology on a simulated dataset, and chromosomal aberrations patterns to classify cancers with plausible results.

References

[1] P. R. Adhikari and J. Hollmén. Patterns from multiresolution 0-1 data. In *UP '10: Proceedings of the ACM SIGKDD Workshop on Useful Patterns*, pages 8–16, New York, NY, USA, 2010. ACM.

[2] P. R. Adhikari and J. Hollmén. Multiresolution Mixture Modeling using Merging of Mixture Components. In S. Hoi and W. Buntine, editors, *Proceedings of 4th Asian Conf. on Machine Learning*, volume 25 of *ACML'12*, pages 17–32. JMLR Workshop and Conference Proceedings, 2012.

[3] M. Baudis. Genomic imbalances in 5918 malignant epithelial tumors: An explorative meta–analysis of chromosomal CGH data. *BMC Cancer*, 7(1):226, December 2007.

[4] R. E. Bellman. *Adaptive control processes - A guided tour.* Princeton University Press, Princeton, U.S.A., 1961.

[5] A. Blum and T. Mitchell. Combining Labeled and Unlabeled Data with Co-training. In *Proceedings of 11th Annual Conf. on Computational Learning Theory*, COLT' 98, pages 92–100, New York, NY, USA, 1998. ACM.

[6] C-W. Hsu, C-C. Chang, and C-J. Lin. A Practical Guide to Support Vector Classification. Technical report, Dept. of Computer Science, National Taiwan University, 2003.

[7] A. Kyriakopoulou and T. Kalamboukis. Clustering as a prior step to classification: an empirical study. *Intl. Journal on Artificial Intelligence Tools*, 20(03):531–548, 2011.

[8] E. R. Mardis. A decade's perspective on DNA sequencing technology. *Nature*, 470(7333):198–203, February 2011.

[9] S. Myllykangas, J. Himberg, T. Böhling, B. Nagy, J. Hollmén, and S. Knuutila. DNA copy number amplification profiling of human neoplasms. *Oncogene*, 25(55):7324–7332, November 2006.

[10] S. J. Pan and Q. Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[11] L. G. Shaffer and N. Tommerup. *ISCN 2005: An Intl. System for Human Cytogenetic Nomenclature(2005) Recommendations of the Intl. Standing Committee on Human Cytogenetic Nomenclature.* Karger, 2005.

[12] P. Smyth. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 10(1):63–72, 2000.

[13] A. S. Willsky. Multiresolution Markov models for signal and image processing. *Proceedings of the IEEE*, 90(8):1396–1458, August 2002.