← Winners' notes. Brian Jones on Incremental Transductive Ridge Regression

Winners' notes. CNSlab team on music instruments recognition →

Winners' notes. Using Multi-Resolution Clustering for Music Genre Identification

APRIL 12, 2011 BY MARCIN WOJNARSKI LEAVE A COMMENT

By Amanda Schierz, Marcin Budka and Edward Apeh (*domcastro*, *BeYou*) from Bournemouth University, UK, 1st and 2nd in Music Genres track of ISMIS 2011 Contest: Music Information Retrieval.

Thanks for this competition – it was great fun. Software used: R, Weka, LibSVM, Matlab, Excel. This was the 2nd competition I had entered (the first being the SIAM biological one) and I only really entered because I had so much undergraduate marking to do! We developed a novel approach to the problem which involved multi-resolution clustering and Error Correcting Output Coding. Our 2nd place approach involved transforming the cluster labels into feature vectors.

Method and Journey:

1. We trained on 50% of the training data using Weka and built an ensemble of a cost-sensitive random forest (number of trees 100, number of features 25), a Bayes Net and a neural network. This resulted in 77.44% on the preliminary dataset. It was very frustrating as we couldn't improve on this. We then looked at semi-iterative relabeling schemes such as Error Correcting Output Coding (using Matlab and LibSVM). This resulted in 81.59% prediction accuracy.

2. We then decided to look at the "statistics" of number of performers, segments, genres etc. We used R to normalize the data (training and test data) and to carry out K-means clustering, k =6 for genres, k=60 for performers, k=2000 for possible songs etc. Taking each set of clusters independently didn't give any information. However, as we had pasted the results into the same file, we noticed a distinct pattern when the cluster results were looked at together – even though no crisp clusters were identified, we noticed that if a training instance was of a different genre from the rest of the cluster then it usually belonged to a different lower granularity cluster. We then built lots of cluster sets for the data (multi-resolution clustering). K was set to 6, 15, 20, 60, 300, 400, 600, 800, 900, 1050, 1200, 2000, 3000, 3200, 5000 and 7000 clusters. At the finest granularity cluster (k=7000) a majority cluster vote was taken using the training instance labels and the test set predictions – the whole cluster was relabelled to the "heaviest" class. If a cluster could not be converged at the finest k-level then we "fell back" to a lower granularity cluster (k=5000) and so on. These new predictions were fed back to the ECOC system and the process was repeated.

3. Figure below shows the overall approach we came up with:

@ TunedIT.org
TunedIT Home
TunedIT Challenges
TunedIT Research
TunedIT Services

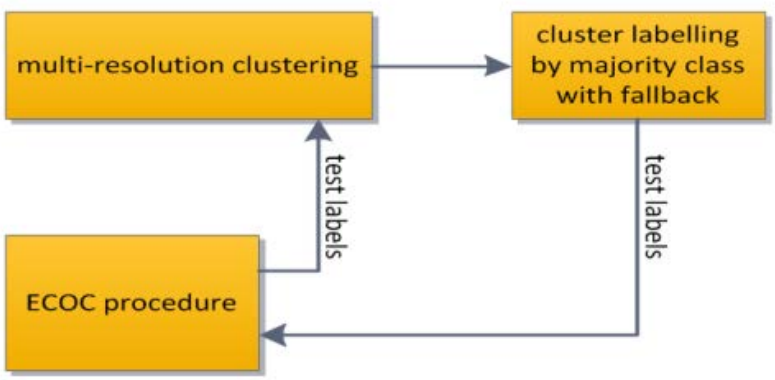
Categories
Select Category

Recent Posts
Winners' notes. CNSlab team on music instruments recognition
Winners' notes. Using Multi-Resolution Clustering for Music Genre Identification
Winners' notes. Brian Jones on Incremental Transductive Ridge Regression
Ed Ramsden on his winning solution in SIAM SDM'11 Contest
Winner's notes. Yuchun Tang on noise deduction to improve classification accuracy in SIAM SDM'11 Contest

Archives
April 2011
February 2011
November 2010
October 2010
July 2010
June 2010
May 2010

RSS
Subscribe in English
Subskrybuj po Polsku

Twitter / TunedIT



4. This was the winning solution and resulted in 0.87507 score on the final test set. For the 2nd place solution, we decided to look at using the cluster assignation labels as feature vectors. This transformed the problem from the original 171-dimensional input space, into a new 16-dimensional space, where each attribute was an identifier of the cluster at one of the 16 levels. So, for example, if instance #7 have fallen into the 3rd out of 6 clusters at the first granularity level, 10th out of 15 clusters at the second granularity level and so on, in the transformed space it would be described as a 16-diemensional vector: [3, 10, …]. Note, that these attributes are now categorical, with up to 7000 distinct values at the highest granularity level. This has limited the number of classifiers we could use.

Our classification system consisted of:
1. Random forest of 1000 unpruned C4.5 decision trees
2. Boosted ensemble of 10 C5.0 decision trees
3. Cross-trained ensemble of 100 Naive Bayes classifiers, trained on different subsets of attributes, each time selected using the Floating Forward Feature Selection method.

We have used majority voting to combine the decisions of these 3 ensembles. After labeling the test dataset using the method described above, we have fed both training and test dataset (this time with the labels from the previous step) to the ECOC system to obtain final predictions. This resulted in 0.87270 on the final test set.

– Amanda Schierz, Marcin Budka, Edward Apeh

FILED UNDER IN ENGLISH

About Marcin Wojnarski

Data scientist. Researcher and practitioner in machine learning, data mining and artificial intelligence since 1996. Founder and CEO of TunedIT.

Leave a Reply

Your email address will not be published. Required fields are marked *

Name *

TunedIT: ECML/PKDD Discovery Challenge 2011: VideoLectures.Net recommender system http://t.co/UxUMBsC via @kdnuggets

TunedIT: New #datamining contest: ECML/PKDD Discovery Challenge 2011: VL.Net recommender, €5500 prizes - http://tunedit.org/challenge/VLNetChallenge

TunedIT: Using Multi-Resolution Clustering for Music Genre Identification: http://t.co/T6NZ2WZ

Blogroll

- Circle of complexity
- JT on EDM
- R-statistics
- Revolutions
- Stochastix

Email *

Website

Comment

You may use these HTML tags and attributes: `<a href="" title=""> <abbr title=""> <acronym title=""> <b> <blockquote cite=""> <cite> <code> <pre> <del datetime=""> <em> <i> <q cite=""> <strike> <strong>`

Post Comment

Notify me of follow-up comments via email.

Subscribe to this site by email

Meta

Register

Log in

Entries RSS

Comments RSS

WordPress.com

Email Subscription

Enter your email address to subscribe to this blog and receive notifications of new posts by email.

Sign me up!