Interpretable Music Categorisation Based on Fuzzy Rules and High-Level Audio Features

Igor Vatolkin and Günter Rudolph

Abstract Music classification helps to manage song collections, recommend new music, or understand properties of genres and substyles. Until now, the corresponding approaches are mostly based on less interpretable low-level characteristics of the audio signal, or on metadata, which are not always available and require high efforts for filtering the relevant information. A listener-friendly approach may rather benefit from high-level and meaningful characteristics. Therefore, we have designed a set of high-level audio features, which is capable to replace the baseline low-level feature set without a significant decrease of classification performance. However, many common classification methods change the original feature dimensions or create complex models with lower interpretability. The advantage of the fuzzy classification is that it describes the properties of music categories in an intuitive, natural way. In this work, we explore the ability of a simple fuzzy classifier based on high-level features to predict six music genres and eight styles from our previous studies.

1 Towards a Higher Interpretability of Classification Models

Recognition of high-level music categories such as genres and styles provides an efficient and automatic way to organise large music collections and recommend new songs. A large number of past and recent studies have addressed this task: Sturm (2012) lists almost 500 references related to genre recognition. The development of new features and complex classification techniques led to significant improvements in the quality of music classification systems. However, in many cases user-centred evaluation criteria as discussed in Hu and Liu (2010) remain completely untouched and are not integrated into the optimisation of parameter settings. One of these criteria is the interpretability of classification models: if some rules are trained to predict music categories, it may be useful for both music scientists and listeners to better understand and interpret their properties.

I. Vatolkin (✉) • G. Rudolph
TU Dortmund, Chair of Algorithm Engineering, 44227 Dortmund, Germany
e-mail: igor.vatolkin@tu-dortmund.de; guenter.rudolph@tu-dortmund.de

B. Lausen et al. (eds.), *Data Science, Learning by Latent Structures, and Knowledge Discovery*, Studies in Classification, Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-662-44983-7_37

A basic chain of algorithms for classification consists of three steps: feature extraction, feature processing, and the training of classification models. Each of these steps may be designed either to facilitate highly comprehensible models as the final output or ignore the request for interpretability.

The first step towards an interpretable classification model is to start with a set of high-level features which are related to music theory and are understood by a music listener rather than by a signal processing expert. The difference between several levels of interpretability of features from the perspective of a listener is very well illustrated in Celma and Serra (2008): They distinguish between three abstraction layers: low-level features which describe the audio signal and physical properties of the sound, mid-level features which correspond to musical characteristics, such as key and mode, and high-level features which are very close to a listener: moods, opinions, memories, etc.

The goal of feature processing is at first to prepare feature vectors for classification, but also to reduce the dimensionality of the original feature matrix, because it may be very large, in particular for short-frame features. Very popular statistical feature processing methods such as principal component analysis are especially dangerous for the keeping of the interpretability because they transform the original feature space (Essid et al. 2006). A suitable solution to strongly reduce the feature matrix keeping the original feature space is to apply selection both on time and feature dimension: to select a limited amount of time frames according to events related to music structure, such as onsets, beats, and tatums (Vatolkin et al. 2012), and apply feature selection for the identification of the most relevant features (Guyon et al. 2006).

The final step is to train classification models. Again, many methods, such as well-established support vector machines, estimate linear combinations of original features or even transform them to higher dimensional spaces. Other successful methods combine the results of many classifiers, e.g., by stacking as proposed in Wolpert (1992) or building ensembles (Zhou 2012). These approaches often lead to a high classification quality, but the models are not comprehensible any more. One of the possibilities to address interpretability is to build classification rules with linguistic variables using fuzzy controllers (Zhang and Liu 2006), or optimise fuzzy controllers with genetic algorithms (Geyer-Schulz 1998). Fuzzy classification was recommended in Mckay and Fujinaga (2006) as the method which "would significantly improve the quality of ground truth, and would make the evaluation of systems more realistic" but still plays a minor role in most music classification applications. In particular, we are not aware of any work which applies fuzzy classification based on a large set of high-level audio features. To name a few related publications, in Friberg (2005) the prediction of emotional expressions was done using fuzzy modelling of so-called cues (tempo, sound level, and articulation). Application of a fuzzy classifier to predict emotions was reported in Yang et al. (2006). In Fernández and Chávez (2012), a fuzzy-rule based system is optimised with the help of evolutionary algorithms to distinguish between classical and jazz recordings using several features which describe frequencies with the strongest amplitudes, and Abeßer et al. (2009) introduced a rule-based framework for genre classification.

2 High-Level Audio Features

In this study we concentrate on audio features, which can be extracted independently of the popularity of songs or the availability of the score. Even if the estimation of some complex signal characteristics is time intensive, this can be done offline only once for the building of the feature set. However, the main challenge is that it is very hard to robustly extract high-level features from the polyphonic signal. This task can be solved to a certain level by machine learning approaches. Probably the first work which integrated this approach for the extraction of high-level features, so-called anchors, is Pachet and Zils (2003). In Vatolkin (2013), we have proposed a novel optimisation approach called sliding feature selection (SFS), where high-level features are iteratively predicted from other high-level and low-level characteristics, and the building of classification models is optimised by means of multi-objective evolutionary feature selection.

Table 1 lists high-level features, which are used in this work as the basis for further fuzzy recognition of genres and styles, providing some examples (second column) and the overall number of feature dimensions. These features can be roughly distinguished in three groups. The first one contains directly implemented mostly harmonic and short-frame characteristics. The second one corresponds to features derived with the help of an SFS-optimised machine learning approach. The

Table 1 Groups of high-level audio features with examples. Dim.: overall number of dimensions of the corresponding group

Group	Examples	Dim.
Directly implemented features		
Chroma and harmony	Tonal centroid, key, strengths of intervals	129
Chord statistics	Number of different chords in 10 s	5
Tempo, rhythm, and structure	Duration, beats per minute	9
SFS-optimised high-level features		
Instruments and effects	Guitar, piano, effects distortion	48
Singing characteristics	Singing solo rough, singing solo polyphonic	56
Harmony	Major, minor	16
Melody	Melodic range > octave, melody linear	32
Moods	Earnest, energetic	72
Structural complexity		
Chord, harmony, instruments, tempo and rhythm complexity		70

last group of features describes structural changes of several high-level groups as introduced in Mauch and Levy (2011). Because for the first group of short-framed harmony characteristics we have estimated the mean and standard deviation values for larger music intervals, the overall number of dimensions of features is 566.

3 Measuring of Feature Relevance with Linguistic Terms

A preliminary step in the design of a fuzzy classifier is to describe the values of features with linguistic terms. Usually not more than 5–7 terms are used, such as *very low*, *low*, *moderate*, *high*, *very high*. The values of features can then be mapped to a membership function which estimates the relationship grade to a category.

Figure 1 provides an example of a relevant feature (top) and not relevant feature (bottom) for the prediction of the category Pop. The segments of songs of the training set which are used as classification instances are marked with small circles. Because in our scenario the songs are assigned as either belonging to a category (positive examples) or not belonging to it (negative examples), the corresponding membership functions are equal to 0 resp. 1. In the upper subfigure it can be observed that songs which belong to the category Pop have the values of the feature "Drum recognition share" always equal or greater than 0.4. On the other side, songs

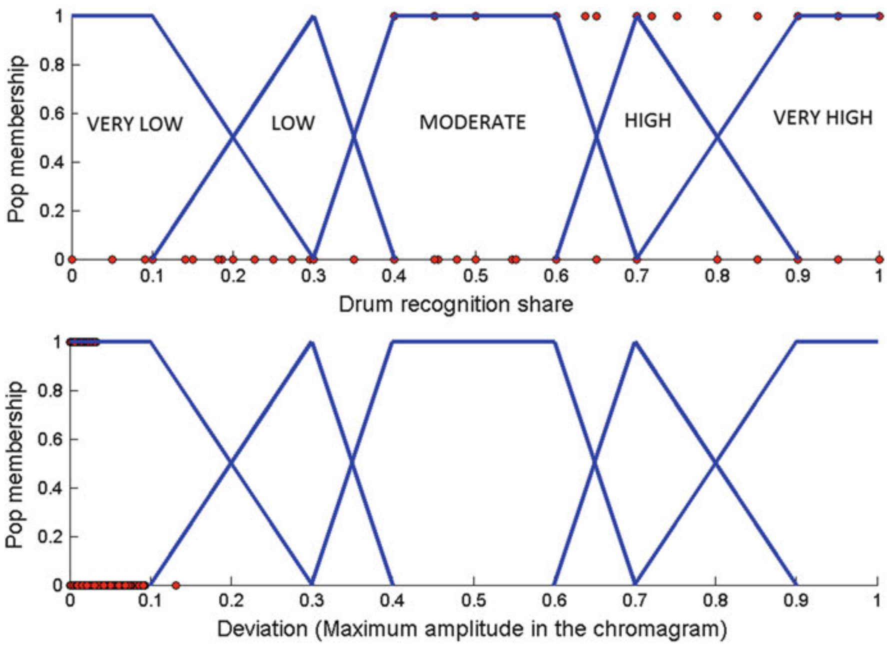

Fig. 1 Features with high (*top subfigure*) and low (*bottom subfigure*) relevance

which do not belong to Pop may contain very different shares of drums between 0 and 1. In other words, if the share of drums is below 0.4 in a song, it is very probable that this song does not belong to the category Pop—at least according to the previously selected songs for the training set.

Figure 1, lower subfigure, provides an example of a feature which is not well suited for the recognition of Pop songs. The deviation of the maximum amplitude of the chromagram is almost always below 0.1 for both positive and negative training songs.

The estimation of individual relevance of features may be used for the prediction of the membership function. If C_k would denote the kth category to predict (e.g., Pop or Classic), X_i would denote the feature i (e.g., share of drum recognitions), and T_j would describe a linguistic term (*very low*, *low*, *moderate*, *high*, *very high*), the membership grade can be calculated using the Bayes theorem as the conditional probability:

$$P(C_k|\text{feature } X_i \text{ is } T_j) = \frac{P(\text{feature } X_i \text{ is } T_j|C_k) \cdot P(C_k)}{P(\text{feature } X_i \text{ is } T_j)}, \tag{1}$$

where the terms in the right half of the equation are estimated from the training data.

Because we use approximately the same number of positive and negative songs in training sets for a better balance, $P(C_k) \approx 0.5$. The non-relevant feature from the bottom subfigure of Fig. 1 outlines a problematic issue of the application of the Eq. (1). For T_j = *very low*, $P(\text{feature } X_i \text{ is } T_j) = 1$ (this feature has always *very low* values). Further, $P(\text{feature } X_i \text{ is } T_j|C_k) = 1$, so that $P(C_k|\text{feature } X_i \text{ is } T_j) \approx 0.5$. However, it is better to set a significantly lower value for the relevance of the rule "IF deviation of the maximum amplitude of the chromagram is *very low* THEN Pop". Therefore, the features which almost always belong to a certain linguistic term may be penalised using the following formula for the estimation of the relevance of a rule "IF feature X_i is T_j THEN category C_k":

$$R(C_k, X_i, T_j) = P(\text{feature } X_i \text{ is } T_j|C_k) \cdot (1 - P(\text{feature } X_i \text{ is } T_j)). \tag{2}$$

Table 2 lists the most relevant rules for the prediction of three music genres (Classic, Pop, Rap) and music styles (ClubDance, HeavyMetal, ProgRock). For simplicity reasons, we omit some details of the feature estimation, such as the underlying supervised classification method. The features in these rules provide a comprehensible description of the properties of the tested categories, compared to low-level characteristics of the audio signal. The linguistic terms *very high* and *very low* belong to rules with highest relevance values. For example, the rule "IF structural complexity of harmony is *moderate* THEN Classic" has the position 383 ($R(C_k, X_i, T_j) = 0.1441$) in the list of rules sorted according to their relevance, and "IF structural complexity of harmony is *high* THEN Classic" has the position 1,251 ($R(C_k, X_i, T_j) = 0.0401$).

Table 2 The most relevant rules for the recognition of three genres and three styles

Rule	$R(C_k, X_i, T_j)$
Genre classic	
IF structural complexity of harmony is *very high* THEN Classic	0.4030
IF melodic range greater than octave is *very high* THEN Classic	0.3821
IF mood Earnest is *very high* is THEN Classic	0.3816
IF mood Stylish is *very low* is THEN Classic	0.3813
Genre pop	
IF singing solo rough is *very high* THEN Pop	0.3844
IF key major is *very low* THEN Pop	0.3498
IF key minor is *very high* THEN Pop	0.3342
IF number of segment changes is *very high* THEN Pop	0.3277
Genre Rap	
IF mood PartyCelebratory is *very high* THEN Rap	0.4895
IF melodic range less than octave is *very high* THEN Rap	0.4699
IF mood Sentimental is *very low* THEN Rap	0.4502
IF singing solo position medium is *very high* THEN Rap	0.4225
Style ClubDance	
IF mood Energetic is *very high* THEN ClubDance	0.3421
IF mood PartyCelebratory is *very high* THEN ClubDance	0.3420
IF melodic range greater than octave is *very low* THEN ClubDance	0.3348
IF singing solo clear is *very high* THEN ClubDance	0.3288
Style HeavyMetal	
IF mood Aggressive is *very high* THEN HeavyMetal	0.3861
IF effects distortion is *very high* THEN HeavyMetal	0.3850
IF singing solo rough is *very high* THEN HeavyMetal	0.3514
IF singing solo clear is *very low* THEN HeavyMetal	0.3505
Style ProgRock	
IF singing solo rough is *very high* THEN ProgRock	0.3142
IF mood Stylish is *very low* THEN ProgRock	0.2910
IF number of segment changes is *very high* THEN ProgRock	0.2904
IF melodic range greater than octave is *very high* THEN ProgRock	0.2867

A further possibility which was not examined for this paper but is promising for future studies is the combination of rules using fuzzy operators for "and" and "or", e.g. "IF structural complexity of harmony is *very high* AND structural complexity of harmony is *high* THEN Classic". However, the number of possible rules to analyse may explode: for simple rules based on a single feature the number of possible rules is already 2,830 (five linguistic terms for 566 high-level audio features). Not only the combination of two and more rules may be helpful for fuzzy classification, but also the optimisation of the definition areas of the linguistic terms, as done in Fernández and Chávez (2012) with the help of evolutionary algorithms.

4 Application of a Simple Fuzzy Classifier

A simple multi-class fuzzy classifier may estimate the average relevance of M most relevant rules and select the genre with a highest value:

$$\hat{C}_k = \max_{k \in \{1,\ldots,C\}} \left(\frac{1}{M} \sum_{m=1}^{M} R\left(C_k, X(m), T(m)\right) \right), \tag{3}$$

where C is the number of (exclusive) genres to predict, $X(m)$ is the feature from the mth rule for the genre C_k, $T(m)$ is the linguistic term from the mth rule for the genre C_k, and the rules are strictly ordered by decreasing $R(C_k, X_i, T_j)$ as defined in Eq. (2). If there are several equal maximum values, ties are broken at random.

The basic challenge of this method is to find the optimal value for M. The one extreme is to apply only the most relevant rule. Because only one high-level feature is used for the prediction of the category in that case, the classification quality may be often too low. For example, we compared the most relevant rules for six genres (Classic, Electronic, Jazz, Pop, Rap, R'n'B) for the identification of the genre of the song "Let Me Put My Love Into You" from AC/DC. The rule with the highest relationship grade classifies this song to R'n'B, the next one to Jazz, and the third one to Pop. However, if we average $R(C_k, X_i, T_j)$ for 50 rules as described in Eq. (3), the song is correctly predicted as belonging to the category Pop.

Another extreme is to apply a very large number of rules. In that case the classification performance can be significantly increased, as illustrated in Fig. 2. Here, the averaged $R(C_k, X_i, T_j)$ of up to the 300 most relevant rules is estimated for the classification of 120 test songs. However, if a high number of rules are used for a genre prediction, too many high-level music features contribute to the final decision and the interpretability decreases. A compromise solution would be to

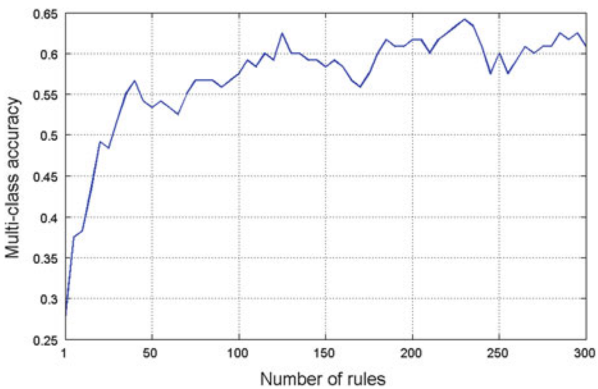

Fig. 2 Multi-label accuracy for the prediction of six music genres using 1–300 most relevant fuzzy rules (with the step size of five rules)

apply 20 rules: this method has the multi-class accuracy of 49.17 %, which has a potential to be improved, but is already significantly above the performance of a random classifier for six genres which would achieve an expected probability of 16.67 %. Another local optimum has the accuracy of 56.67 % (40 rules).

The aggregation of rules may also be applied for binary classification. Here we compare the results to Vatolkin (2013), where the same set of high-level features was used, however four different supervised classification methods and evolutionary multi-objective feature selection were applied for the recognition of genres and styles. In spite of significant improvements of the classification performance, the complex methods developed for the aforementioned study have also some drawbacks: the optimisation requires large computing times (up to more than 24 h if combined with support vector machines), and the interpretability of original high-level features suffers if a complex classification method (support vector machine or ensemble of many decision trees) provides the best classification model. If a simple fuzzy classification model (as discussed above) contains a limited number of rules M, other advantages are that the models have very small storage demands and the classification of new songs can be done very fast.

Table 3 compares the results from both studies. The column "HL-all" lists the classification errors if complete feature set is used for the classification with four tested methods, and the column "HL-FS" describes the errors after the optimisation by means of multi-objective evolutionary feature selection. All models were trained

Table 3 Balanced classification error for experiments with several classification methods (columns HL-all, HL-FS) and fuzzy rules (columns HL-fz10, HL-fz50, HL-fz100)

Task	HL-all	HL-FS	HL-fz10	HL-fz50	HL-fz100
Recognition of genres					
Classic	0.0365	0.0137	0.0524	0.0524	0.0238
Electronic	0.2010	0.1191	0.2571	0.2524	0.2524
Jazz	0.0866	0.0605	0.1904	0.1904	0.1904
Pop	0.2890	0.1270	0.4156	0.3444	0.3244
Rap	0.0852	0.0650	0.1143	0.1143	0.1143
R & B	0.1931	0.1484	0.3000	0.2762	0.2762
Recognition of styles					
AdultContemporary	0.2358	0.1344	0.2909	0.2227	0.2227
AlbumRock	0.2084	0.1066	0.3500	0.3500	0.3500
AlternativePopRock	0.2015	0.1092	0.1875	0.1875	0.1875
ClubDance	0.2484	0.1389	0.2500	0.2500	0.2500
HeavyMetal	0.1384	0.0778	0.1024	0.1024	0.1024
ProgRock	0.1818	0.0973	0.3963	0.3963	0.3963
SoftRock	0.2253	0.1197	0.3003	0.2147	0.2147
Urban	0.1467	0.0837	0.2273	0.2273	0.2273

For details see the text

from small training sets of ten positive and ten negative music pieces for each category. The balanced classification error was estimated for the independent validation set of songs not involved into the learning procedure. For details, see Vatolkin (2013).

The third column ('HL-fz10') contains the smallest errors from the combinations of 1 to 10 most relevant rules after the simple fuzzy classification discussed above (Eq. (3), $M = 1, \ldots, 10$). Similarly, for columns "HL-fz50" and "HL-fz100" the aggregation of up to 50 and up to 100 most relevant rules was estimated. The performance of a simple classifier using up to 100 fuzzy rules depends on the category: e.g., for Electronic and Pop the error is always higher than using four methods and all features (the column "HL-all"), and on the other side for AlternativePopRock and HeavyMetal the aggregation of up to ten rules performs already better than the classification with four methods using all high-level features. In all cases the classification based on fuzzy rules leads to significantly higher errors than the classification with four methods using the optimised feature set ('HL-FS'). However, the fuzzy approach has still enough room for optimisation without losing the interpretability.

5 Summary and Outlook

In this paper we have discussed a basic approach to music classification where each algorithm step is designed to provide as interpretable outputs as possible, from comprehensible high-level audio characteristics to a simple fuzzy classifier, which aggregates a limited number of categorisation rules which describe well the most important musical properties of music genres and styles. The results show that although the classification performance is inferior to the approach where several classification methods and feature selection are applied, they are still significantly better than for a random classifier, the method is very fast, and the classification models are comprehensible for listeners and music scientists.

There exist several possibilities to improve our basic implementation with extended techniques keeping the high interpretability of classification models. In particular, the combination of different high-level features using fuzzy "and" and "or" operators for the corresponding linguistic terms is promising, among others because the features are then not treated independently and may be relevant for a certain category only in their combination. The adaptation of the definition areas of linguistic terms as proposed in Fernández and Chávez (2012) and the application of rule selection similar to feature selection as done in Vatolkin (2013) are other starting points for further investigations.

Acknowledgements We thank the Klaus Tschira Foundation for the financial support.

References

Abeßer, J., Lukashevich, H., Dittmar, C., & Schuller, G. (2009). Genre classification using bass-related high-Level features and playing styles. In *Proceedings of the 10th Int'L Conference on Music Information Retrieval (ISMIR)* (pp. 453–458).

Celma, Ò., & Serra, X. (2008). FOAFing the music: Bridging the semantic gap in music recommendation. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 6*(4), 250–256.

Essid, S., Richard, G., & David, B. (2006). Musical instrument recognition by pairwise classiffication strategies. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(4), 1401–1412.

Geyer-Schulz, A. (1998). Fuzzy genetic algorithms. In H. T. Ngyen & M. Sugeno (Eds.), *Fuzzy systems*. Boston: Kluwer Academic Publishers.

Guyon, I., Nirkavesh, M., Gunn, S., & Zadeh, L. A. (2006). *Feature extraction. foundations and applications*. Berlin/Heidelberg: Springer.

Fernández, F., & Chávez, F. (2012). Fuzzy rule based system ensemble for music genre classification. In *Proceedings of the 1st International Conference on Evolutionary and Biologically Inspired Music, Sound, Art and Design (EvoMUSART)* (pp. 84–95). Berlin: Springer.

Friberg, A. (2005). A fuzzy analyzer of emotional expression in music performance and body motion. In J. Sundberg & B. Brunson (Eds.), *Proceedings of Music and Music Science*.

Hu, X., & Liu, J. (2010). User-centered music information retrieval evaluation. In *Proceedings of the Joint Conference on Digital Libraries (JCDL) Workshop: Music Information Retrieval for the Masses*.

Mauch, M., & Levy, M. (2011). Structural change on multiple time scales as a correlate of musical complexity. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 489–494).

Mckay, C., & Fujinaga, I. (2006). Musical genre classification: Is it worth pursuing and how can it be improved? In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)* (pp. 101–106).

Pachet, F., & Zils, A. (2003). Evolving automatically high-level music descriptors from acoustic signals. In *Proceedings of the 1st International Symposium on Computer Music Modeling and Retrieval (CMMR)* (pp. 42–53).

Sturm, B. (2012). A survey of evaluation in music genre recognition. In *Proceedings of the 10th International Workshop on Adaptive Multimedia Retrieval (AMR)*.

Vatolkin, I., Theimer, W., & Botteck, M. (2012). Partition based feature processing for improved music classification. In W. A. Gaul, A. Geyer-Schulz, L. Schmidt-Thieme, & J. Kunze (Eds.), *Challenges at the interface of data analysis, computer science, and optimization* (pp. 411–419). Berlin: Springer.

Vatolkin, I. (2013). Improving supervised music classification by means of multi-objective evolutionary feature selection. PhD thesis, Department of Computer Science, TU Dortmund, 2013.

Yang, Y. -H., Liu, C. -C., & Chen, H. H. (2006). Music emotion classification: A fuzzy approach. In: K. Nahrstedt, M. Turk, Y. Rui, W. Klas, & K. Mayer-Patel (Eds.), *Proceedings of the 14th ACM International Conference on Multimedia* (pp. 81–84).

Wolpert, D. (1992). Stacked generalization. *Neural Networks, 5*(2), 241–260.

Zhang, H., & Liu, D. (2006). *Fuzzy modeling and fuzzy control*. Boston/Basel/Berlin: Birkhäuser.

Zhou, Z. -H. (2012). *Ensemble methods: Foundations and algorithms*. Boca Raton: CRC Press.