An Automatic Segmentation Method of Popular Music Based on SVM and Self-similarity

Feng Li(✉), You You, Yuqin Lu, and YuQing Pan

School of Computer Science and Communication Engineering,
JiangSu University, XueFu Road No. 301, Zhenjiang 212013, JiangSu, China
fengli@ujs.edu.cn, getyouyou@126.com

Abstract. Popular music segmentation plays an important role in popular music content retrieval, melody extraction and semantic understanding etc. Segmentation boundary detection is one of the key technologies in lots of conventional algorithms. In this paper, we propose an automatic segmentation approach that combines SVM classification and audio self-similarity segmentation, which firstly separate the sung clips and accompaniment clips from pop music by using SVM preliminary classification, and then heuristic rules are used to filter and merge the classification result to determine potential segment boundaries further. Finally, self-similarity detecting algorithm is introduced to refine our segmentation results in the vicinity of potential points. Experiment shows that our approach can achieve more sophisticated and accurate results by using the local change context substantially to determine the segmentation boundaries.

Keywords: Popular music segmentation · SVM · Self-similarity

1 Introduction

Music segmentation partitions the music file into its composite sound clips in time domain according to the melody change information. It plays an important role in music process application fields such as content-based retrieval, melody extraction, music emotion recognition and audio database management etc. The popular music occupies a large proportion in terms of the quantity, which indicates that research on the popular music segmentation technology is meaningful for promoting the development of audio automatic segmentation.

There are lots of previous researches on popular music automatic segmentation has been explored. J.Foote[1, 2] uses the music signal time-domain and frequency-domain features to calculate the similarity measurement and constructs self-similarity[3] metric, then designs checkboard kernel with various scale skillfully to compute the novelty value to evaluate the changes of objective music in a time-lag. In this approach the segmentation boundaries are found by selecting the peaks of novelty series, which are sensitive to the checkboard kernel scale, especially when the kernel size is small, the boundaries will be difficult to determine. Jessie Xin Zhang[4] employs the RMS and SP features to detect the silence clips in the audio file, then uses

Q. Zu et al. (Eds.): HCC 2014, LNCS 8944, pp. 15–25, 2015.
DOI: 10.1007/978-3-319-15554-8_2

these clips to separate the file into different segments and constructs the similarity image for each segment, finally acquires the boundaries by using image edge detect algorithm. In consequence of relaying on silence clips detection, the application scenario of this approach is restricted, it often be applied to process broadcasting and TV audio files to gain the music and speaking partitions. Masataka Gotow[5] takes account of the chroma feature to compute the similarity of music signal frames, the music melody is segmented by finding the repeating sections. This approach gives a good result but needs vast computing cost when finding the repeating section duration. L.Shuang[6] considers the relationship between music lyric text and melody structure, the timestamp information in lyric text is used to give some tough segmentation, then the result is optimized by taking use of some heuristic rules based on the lyric sentences statistic characteristics, this properties includes the characteristic of last word intonation and the word number in each sentence. This approach gives a good performance but its shortage is the essential of the timestamp information in the lyric file.

The main limitations of the methods mentioned above are segmentation boundaries determining difficulties and the limited application scenarios. In this paper, we combine SVM classification algorithm with audio self-similarity detection method to separate the sung and accompaniment clips from music signal innovatively to refine our music automatic segmentation result. Experiments results show that our method has the capability to overcome the limitation of boundaries determining and improve the segmentation performance significantly with acceptable computing cost.

By summarizing our perception experiments of popular music, we can infer that the popular music rhythm changes usually occurred in accompaniment which last a long time and the beginning or the end of every sung line. We call these change points as our potential boundaries. According to these perception experiments we employ the SVM classification algorithm to pre-separate our input popular music signal at first, this procedure will output a discrete label series which correspond to the sung and accompaniment [7, 8] discriminate result of each frame respectively. Then we utilize some heuristic rules to filter and merge our pre-process SVM classification series, which can minimize the misclassification effect brought from the first step, and more accurate potential segmentations will be provided after this process. Finally, we take advantage of the sensitivity of novelty detection algorithm based on self-similarity measurement in the adjacent of our potential points with different scale kernel to gain our final segmentation boundaries. In comparison with the method based on self-similarity deployed to detect the music segmentation boundaries without pre-process procedure, our method use the self-similarity detect algorithm only on the potential change points which carry more local music structure context information rather than the entire music signal, and experiment shows our method gains more accurate segmentation result.

The rest of this article is organized as follows. In section 2, we present the overall flow diagram of the method and introduce the technology details. Experiment results and relevant analysis will be given in section 3. Finally, we conclude this paper with a review and some applications will be given conventionally.

2 Automatic Segmentation Algorithm Details

Fig.1 shows how this method works for our popular music automatic segmentation. In this method, we first need to construct a popular music training set to train our SVM classification model. The training set is composed of vast music signal samples which include both sung and accompaniment sections, the duration of each sample is between 5 to 30 seconds, all of these samples have been labeled manually. For each sample we divide it into signal frames without overlap and extract frequency-domain features as the input of SVM training procedure. After the SVM classification model parameters by using the training set has been derived, we use this model to pre-separate the objective music into sung and accompaniment clips, then these clips will be merge and discard by using some heuristic rules, the final output of this stage is potential points mentioned previously. On this basis, these potential points are processed by the self-similarity detection algorithm to produce the segmentation boundaries at last.

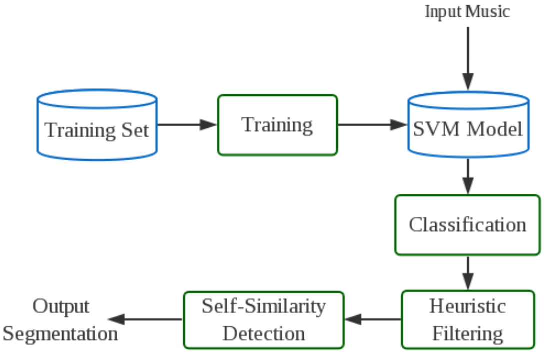

Fig. 1. Popular Music Segmentation

2.1 SVM Preliminary Classification

SVM utilizes kernel function to map the original features to the high-dimensional space which often changes some linear inseparable problem into a linear separable one. SVM has been widely used in many fields and it usually provides higher classification accuracy than other classification algorithms. For the binary classification, the final decision function can be calculated as follows:

$$f(x) = sign(\sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b) \tag{1}$$

For general application with high performance, the common selection of kernel function is Gaussian radial basis kernel (RBF).

$$K(x, z) = \langle \phi(x) \bullet \phi(z) \rangle = \exp(-\frac{\|x - z\|^2}{2\sigma^2}) \tag{2}$$

The discriminate procedure between sung and accompaniment is equivalent to a binary classification problem. In this paper, we select 12-dimensions MFCC[9], its 12-dimensions first-order difference coefficients $\Delta\mathrm{MFCC}$ and 16-dimensions sub-frequency band spectral energy ratio (SFR) as our segmentation feature inputs. The SFR can be used to represent the formant phenomenon which cased by the vocal cord vibration shows in Fig.2. Straightforwardly, we divide the frequency into some sub-frequency bands averagely and calculate the ratio of sub-band spectral energy to total energy. The SFR of qth dimension in pth frame is calculated as follows:

$$SFR[p,q]=\frac{\sum_{i=len*q/\mathrm{n}}^{len*(q+1)/n}\left\|S_p[i]\right\|^2}{\sum_{i=0}^{len}\left\|S_p[i]\right\|^2} \tag{3}$$

The parameter *len* means FFT length, S_p means the spectral of pth frame and n is divide number.

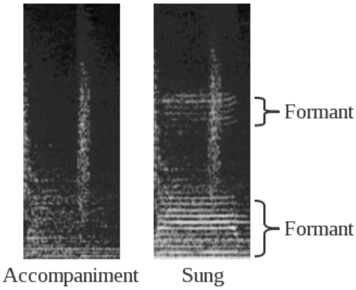

Fig. 2. Spectral formant phenomenon between accompaniment and sung in the spectral with same sampling time

For all samples in the training set, we extract and scale these 40-dimensions features, using the grid parameter search algorithm[10] to find the optimal SVM model. Likewise, we compute these features for all signal frames of our objective music file and use the SVM model to gain our pre-separate classification result, here, accompaniment clips are the negative samples.

2.2 Heuristic Filter and Merge

In the preliminary classification procedure, the segmentation is rough as a result of the misclassification and any other noise effect. This rough classification result is too discrete and crude to our self-similarity detection procedure, we need to merge the small clips and modify the misclassification samples by using some heuristic rules [11]. Through the observation of the music signal, here we propose three filter rules as follows：

1. Music signal changes slowly which indicates the sung and accompaniment cannot change suddenly, so the classification label series are continuous.
2. At the end of sung sentence, the formant phenomenon will be subdued because of the singing rhyme. In this case, the misclassification rate rises and the fluctuation of classification series is obvious.
3. Sometimes the continuous misclassification means this part of the signal is different with the homogeneous segmentation before, and we can add this start time to our potential change points queue.

According to these experiment observation heuristic rules, we can design our filter and merge logics that can be applied on the pre-separate result directly. Rule 1 gives the first filter formula as follows:

$$C[i] = \begin{cases} C[i-1] & \text{if } \mathrm{C[i-}w\mathrm{:i-1] = C[i+1:}i+w]\,\&\,\&C[i]! = C[i-1] \\ C[i] & \text{otherwise} \end{cases} \tag{4}$$

In formula (4), parameter C represents the pre-classification result, w is filter window length. And formula (4) indicates that if the two neighborhoods with w duration has same classification label and the center ith frame has an inverse label, then ith frame should be considered as a misclassification and needs to be modified. Next, we give our second filter and merge policy according to the rule 2 as follows:

(i) Traversing the pre-separate result C, if C[i]=-1, then perform step (ii).
(ii) Start with i, using window with length of w, computing the sum of the absolute value of first-order differential value in this window. The sum is denoted by $\Delta\Sigma$, if $\Delta\Sigma$>w then perform step(iii), otherwise perform step (i).
(iii) Modifying all the labels in the window with length w to -1 and perform step (i).

In order to emphasize the point of this policy, here we give the explanation of step (ii). Considering with the uniform misclassification case, the positive and negative labels appear alternately, then $\Delta\Sigma$ equals to 2w. Correspondingly, $\Delta\Sigma$ equals to 2 when the misclassification is not exist in the compute window. In this paper, we take the discrimination threshold equals to w as the compromise scheme. Now, by using this policy we can detect the parts of singing rhyme in the music easily by observing the series fluctuation.

In contrast with the rule 1, 2 mentioned above, rule 3 provides us a guideline to treat the continuous misclassification rather than an actual filter and merge algorithm or procedures. However, it is found that there exist some conflicts between rule 1 and rule 2 from the point view of practice. In this paper we strongly recommend using rule 2 firstly. The clips with short duration will be merged and the abnormal classification error will be corrected as well after using these filter and merge policies based on heuristic rules listed above. What calls for special attention is that the clips with continuous same labels will be removed except the start and end ones which decide the boundary of the segmentation [12]. So far we have got some real pre-separate result about the sung and accompaniment partitions of the music signal to analysis.

Fig.3 gives an example of the result changes at each stage among the procedure. The song name is ChunXue, and the black areas represent accompaniment partitions and the areas with gray color are sung ones. This demonstration shows that the original SVM classification result can be filtered and merged effectively in this phase.

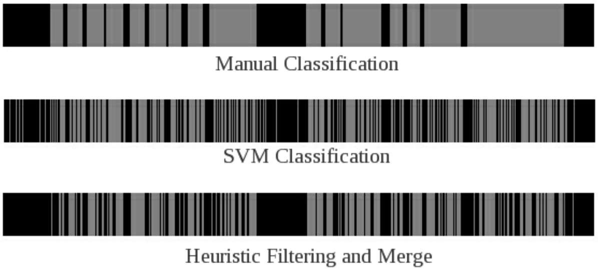

Fig. 3. Sung and Accompaniment separation contrast diagram

2.3 Self-similarity Segmentation Detection

According to the melody structure of popular music, the distribution of perception change points is regular. The potential music rhythm change points can be summarized into two categories.

1. Music signal often changes around the beginning or end of each sung sentence.
2. Music signal often changes in the accompaniment with long duration such as Intro, Bridge and Outro partitions.

Fortunately, we can determine the vast majority of the start and end timestamp of each sung sentences in the music by using the pre-process method mentioned above. Similarly, we can pick out long duration accompaniment partition from the accompaniment recognition results as well. Next, the segment detection method based on self-similarity will be applied at these potential change points.

In the traditional self-similarity detection method[13,14,15,16], frequency-domain features will be extract for each signal frame at first, then cosine similarity will be used to measure the distance between each feature vector, the distance *Dist(i,j)* between *i*th frame and *j*th frame can be compute as follows:

$$Dist(i,j) = \frac{V[i] \bullet V[j]}{\|V[i]\| * \|V[j]\|} \tag{5}$$

V[i] is the feature vector of *i*th frame. All of these distances compose a 2-dimensions similarity metric *S*, then a checkboard metric *C* will be constructed to move along the diagonal of the similarity metric to calculate the novelties which are used for measuring the variation trend of the music structure, and the *i*th novelty is calculated as follows:

$$Novelty(i) = \sum_{m=-L/2}^{L/2} \sum_{n=-L/2}^{L/2} C(m,n)S(i+m,i+n) \tag{6}$$

L is the checkboard kernel size, different size kernel function can be constructed by using Kronecker product:

$$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix} \tag{7}$$

There are some problems if we use this algorithm for the input music file directly. Using small size kernel makes algorithm too sensitive to detect the change points, which means the quantity of segmentation will be large, and the decision of threshold is difficult even if we construct a fixed size maximum heap to select the segmentation boundaries. However, if we use large size kernel the performance will be reduced but the results will be rough. In this paper, we use this detection algorithm for all the potential change points instead of entire music signal. These potential points carry more local melody context information which helps us determine our segmentation boundaries easily and refine the final result. Considering with the first category change points, we apply the algorithm in the fragments which center with the end timestamp of each sung sentence like this:

(i) Traversing all the rough segmentations with positive labels.
(ii) Intercepting a fragment which center with the end timestamp of the sung segmentation, setting the fragment length equal to w, using self-similarity detect algorithm with small size kernel and calculate the novelties.
(iii) Update the segmentation end boundary equal to the time when the novelty has maximum value.

For another category which includes accompaniment segmentations with long duration such as Intro, Bridge and Outro partitions etc. In compare with process of first category, here we use large size kernel to detect all the concrete change points as follows:

(i) Traversing all the rough segmentations with negative labels.
(ii) Using large size kernel to calculate the novelties among the segmentation.
(iii) Building maximum heap to select fixed number of novelties peaks. Removing the original segmentation boundary, and add the timestamps corresponding to these peaks as the refine segmentation result.

3 Experiment

In order to train and test the proposed method in this paper, 100 Chinese and 100 English popular music segmentations have been captured and labeled manually to build our training music set. This set covers a wide range of popular music genre (Rock, Blues, Jazz, Country, Folk, etc.). More specifically, all of these fragments are sampled with the frequency of 22.05 kHz. As the first step, we employ hamming window function to split our samples into signal frames without overlap, and for convenient operation the length of each frame is 46ms so that there are about 1024 sample points in each frame. Then the 40-dimensions frequency-domain features are extracted as the training inputs. After using the grid parameter search algorithm we get our optimal SVM [17] model with 88.67% cross validation accuracy, where parameter *C* equals to 512 and *gamma* is 0.125. 5 Chinese and 5 English songs have been selected to test our model and algorithm. The window length in the filtering phase derived by rule 1 is equal to 7 and in the merge stage the value is 20. The pre-separate results are measured by Classification Recall Rate (CRR) which can be calculated as follows:

$$CRR = \frac{Correct\ Classification\ Frames}{Total\ Frames} \tag{8}$$

In our experiment, the average Recall rate is 83.74%. It is found that the classification accuracy will be improved slightly after using rule 1. However, for the rule 1, we tend to utilize its ability to reduce the negative effects brought by noise and misclassification in pre-process results. Rule 2 is capable of handling, by contrast, the continuous misclassification case and capturing more potential points even if the classification accuracy may decline in some situations.

Finally, we use the segmentation algorithm based on self-similarity for each potential change point. For the first category case whose change points appear at the beginning or end of each sung sentence, kernel size is equal to 16 and window length is 60 with 3s process duration. The novelties will be messy if we use the small kernel for the long duration accompaniments, we chose 64-size large kernel to relieve the dilemma of boundary determine. Fig.4 shows the procedure of using self-similarity segmentation detection algorithm on the entire music signal of ChunXue without pre-process. It clearly shows that the selection of local segmentations boundaries is subject to the global threshold value. The boundaries could not associate with the local rhythms contexts and the segmentation result is rough in this dilemma.

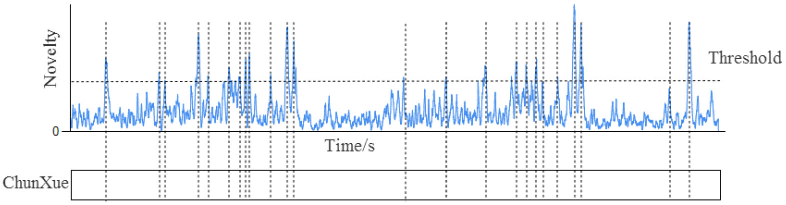

Fig. 4. Self-similarity segmentation detect algorithm without pre-process

By contrast, Fig.5 demonstrates the segment actions on same music signal with our algorithm, and the segmentation object has been pre-classified into sung and accompaniment elements so that the potential points can be figured out by the filter and merge operations, then the self-similarity segmentation detection algorithm has been employed on each point rather than the entire music signal. The duration of the Intro segmentation of the example music is 21s, 4 detail segmentation change boundaries have been found by using our algorithm, and the locations of these boundaries are 8s, 14s, 20s that corresponding to the actual melody change timestamps respectively. At the start of the Bridge segmentation, more precise boundary can be updated by self-similarity detects method with small size checkboard kernel, in the experiment ,the start of the Bridge partition is at 2'0'' and the end of adjacent sung sentence is at 1'57'', therefore, this small duration accompaniment can be merge into the sung sentence in practice application.

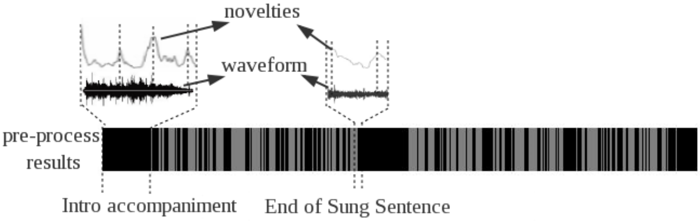

Fig. 5. Self-similarity segmentation detect algorithm at potential points with pre-process

In order to compare the final segmentation results directly, we have used both the entire self-similarity segmentation detect algorithm without pre-process and the modified method proposed in this paper to segment our test music files at the same time. Manual segmentation based on human perception results are also provided as a reference. To quantize the experiment result intuitively, we use segmentation precise rate (SPR) as the measure value, the SPR can be calculated as follows:

$$SPR = \frac{Auto\ Correct\ Segmentations}{Total\ Manual\ Segmentations} \times 100 \tag{9}$$

The definition of auto correct segmentation in this formula is: For each segmentation boundary in the final results, if there exists a boundary makes the difference of them less than a minor term ε hold up, then this segmentation boundary is correct. In our experiment ε takes 0.5s. In Table 1 the final precise result of these methods for 10 test music files is presented. Symbol MS is manual segmentations and PSR is our results precise rate, TSR is the result of the traditional method based on self-similarity without pre-process. It can be observed that our method with SVM pre-classification process gets a significant boost in performance, and the difficult of segmentation boundaries determine can be overcome partly.

Table 1. Performance comparison on 10 test music files include both Chinese and English songs

Song	MS	TSR	PSR	Song	MS	TSR	PSR
ZuiChuDeMengXiang	72	70.5	83.4	Let it go	51	80.9	83.3
QieLangYu	56	71.7	79.1	Devotion	50	74.6	78.8
ChunXue	65	73.7	82.9	Dying in the sun	60	81.8	84.6
PiaoXue	67	81.3	86.6	Moonlight Shadow	61	78.8	82.4
GuangHuiSuiYue	76	69.6	78.3	Heartbeats	45	78.9	83.5

4 Conclusion

In this paper, we take advantage of sung and accompaniment structure information of popular music to assist our automatic segmentation algorithm based on self-similarity to refine the final segmentation result. We use MFCC, $\Delta\mathrm{MFCC}$ and sub-frequency band spectral energy ratio (SFR) as our music signal frame descriptors to train a binary SVM classification model. In the pre-process phase, SVM classification model is used to separate the sung and accompaniment frames. As a result of misclassification and noise, some heuristic filter and merge rules or guideline have been proposed to optimize the rough discrete pre-separate segmentations. After these procedures, the potential music change points can be found, and then the self-similarity segmentation algorithm with the input of potential points is used to refine segmentation boundaries invariably in different category. Experiment result shows that the method has a significant improvement on the accuracy of boundaries determine and raise performance up to 13.4%. The experiment from our practice engineering project shows the output of this method can be applied in the digital fountain and stage lighting control fields excellently.

References

1. Foote, J.: Automatic audio segmentation using a measure of audio novelty. In: Proceedings of IEEE-ICME, vol. I, pp. 452–455 (2000)
2. Cooper, M., Foote, J.: Summarizing popular music via structural similarity analysis. In: 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 127–130 (2003)
3. Foote, J.T., Cooper, M.L.: Media segmentation using self-similarity decomposition. In: Proc. SPIE Storage and Retrieval for Multimedia Databases, 2003, vol. 5021, pp. 167–175 (2003)
4. Zhang, J.X.: A two phase method for general audio segmentation. In: Multimedia and Expo, 2009, pp. 626–629 (2009)
5. Goto, M.: A chorus-section detecting method for musical audio signals. In: Proc. ICASSP, 2003, vol. V, pp. 437–440 (2003)
6. Shuang, L.: Lyrics-based music structure analysis of Chinese pop song. In: NCMT2009, pp. 86–87 (2009)

7. Berenzwig, A.L.: Locating singing voice segments within music signals. In: 2001 IEEE Workshop on Application of Signal Processing to Audio and Acoustics, pp. 119–122 (2001)
8. Fujihara, H.: F0 Estimation method for singing voice in polyphonic audio signal based on statistical vocal model and viterbi search. In: Proceedings of Acoustics, Speech and Signal Processing, 2006. ICASSP 2006, vol. 5, pp. 253–256 (2006)
9. Peeters, G.: A large set of audio features for sound description (similarity and classification) in the CUIDADO project. CUIDADO Project Report (2004)
10. Shunjie, H., Qubo, C., Meng, H.: Parameter selection in SVM with RBF kernel function. In: IEEE World Automation Congress (WAC), 2012, pp. 1–4 (2012)
11. Chang, C.-W.: A heuristic approach for music segmentation. In: Innovative Computing, Information and Control, pp. 228–232 (2007)
12. Peiszer, E., Lidy, T.: Automatic audio segmentation: segment boundary and structure detection in popular music. In: The 2nd International Workshop on Learning Semantics of Audio Signals, LSAS 2008, pp. 48–59 (2008)
13. Scarfe, T., Koolen, W.M.: A long-range self-similarity approach to segmenting DJ mixed music streams. In: Artificial Intelligence Applications and Innovations IFIP Advances in Information and Communication Technology, vol. 412, pp. 235–244 (2013)
14. Wang, H., Xu, Y., Li, M.: Study on the MFCC similarity-based voice activity detection algorithm. In: Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on Artificial Intelligence, pp. 1391–4394 (2011)
15. Wu, Q., Zhang, X., Lv, P.: Perceptual similarity between audio clips and feature selection for its measurement. In: Chinese Spoken Language Processing (ISCSLP), 2012 8th International Symposium on Chinese Spoken Language Processing, pp. 387–391 (2012)
16. Xu, C., Maddage, N.C., Kankanhalli, M.S.: Automatic Structure Detection for Popular Music. IEEE Multimedia **13**(1), 65–77 (2006)
17. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27 (2011). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Springer

http://www.springer.com/978-3-319-15553-1

Human Centered Computing
First International Conference, HCC 2014, Phnom Penh,
Cambodia, November 27-29, 2014, Revised Selected
Papers
Zu, Q.; Hu, B.; Gu, N.; Seng, S. (Eds.)
2015, XIX, 913 p. 406 illus., Softcover
ISBN: 978-3-319-15553-1