2010 Fourth International Conference on Digital Society

Customer Segmentation Architecture Based on Clustering Techniques

Guillem Lefait
University College Dublin,
School of Computer Science and Informatics,
Belfield, Dublin 4, Ireland.
guillem.lefait@ucd.ie

Tahar Kechadi
University College Dublin,
School of Computer Science and Informatics,
Belfield, Dublin 4, Ireland.
tahar.kechadi@ucd.ie

***Abstract*—Knowledge on consumer habits is essential for companies to keep customers satisfied and to provide them personalised services. We present a data mining architecture based on clustering techniques to help experts to segment customer based on their purchase behaviours. In this architecture, diverse segmentation models are automatically generated and evaluated with multiple quality measures. Some of these models were selected for given quality scores. Finally, the segments are compared. This paper presents experimental results on a real-world data set of 10000 customers over 60 weeks for 6 products. These experiments show that the models identified are useful and that the exploration of these models to discover interesting trends is facilitated by the use of our architecture.**

I. Introduction

Knowledge on consumer behaviours is essential to provide them high-quality services and to keep them loyal. However, services towards customers have evolved with the society computerisation. First, on a global market, the number of effective customers has increased dramatically. Second, the customisation possibility of products together with the dematerialisation of services have led to an important growth in the number of available products in stores. Third, customers are looking for instant consumption : the product has to be available immediately. Finally, customers are more volatile than before, they will churn if they are not satisfied immediately or if another company offers better services. All those aspects existed in the past, but today, retailers face a scalability issue both in terms of number of customers to satisfy and products to sell. Therefore, companies need automatic methods to summarise and to highlight trends and irregularities.

Clustering, consists of creating groups, such as objects inside same groups are very similar and objects of different groups are very dissimilar [1]. Clustering results are used for different objectives : to explore the data set, to condense it into a small set of representative points, to organise the data and for decision making.

Segmentation is the ability to recognise groups of customers who share the same, or similar, needs [2]. Not all the customers in a broadly-defined market have the same needs, therefore the segmentation enables companies to provide specific products or services to different customers. The use of clustering to automatically provide a segmentation is not recent and has been done for two main objectives : 1) to identify groups of entities that share certain common characteristics and 2) to better understand consumer behaviours by identifying homogeneous groups [3]. However, there are different challenges when using clustering to perform the segmentation : which data to select, how many clusters to produce and how to evaluate the clustering results.

The selection of the appropriate data to partition the customers into groups has two main challenges. The first challenge is to formalise implicit data. The RFM values that represent respectively the **R**ecency, the **F**requency and the **M**onetary score of a customer are an example of derived information and is used in [4] and [5] to segment bank customers. A second challenge is to select a relevant subset of the available features to perform the clustering. Data on consumers may be divided into two main groups : the *a priori* knowledge (the demographic information, such as the age or the gender) or the post hoc knowledge (the information derived from the purchase behaviour). Dennis [6] shows that the behaviour information are more correlated with the spending than the descriptives variables. Moreover, the selected features may have an impact on the consistency of the segmentation [7]. An identified segment based on *post hoc* data might remain essentially the same over time, but the demographic characteristics of that segment might change.

The K-Means clustering algorithm has been often selected [8], [9] to make the segmentation because of its simplicity and its efficiency.

Very few solutions have been proposed to evaluate the quality of the customer segmentation. Manual investigation is often the solution used to assess the relevance of the clusters [4], [5], [8]. In [10], the segmentation is evaluated by the accuracy to predict loyalty of unknown customers. In [11], sales forecasting is used on the segmentation result. The segmentation usefulness is then directly correlated with the result of the process that will take advantage of the clustering. Similarly, segmentation is assessed in [12] by the predictive gain. Authors provided an algorithm, Direct Grouping, that tries to merge iteratively two sub-clusters if the resulting cluster lead to an improvement in the accuracy of the predictor.

978-0-7695-3953-9/10 $26.00 © 2010 IEEE
DOI 10.1109/ICDS.2010.47

Considering the different difficulties that come across in the production of customer segmentations, we propose an automatic architecture to search and to select customer segmentation without manual investigation.

An architecture with the objective to identify hot spots is described in [13]. First, clusters are generated and cluster descriptions are extract. Then the interest of each cluster is defined by the domain user and finally the interest criteria are evolved given expert feedbacks.

Our approach differs in the following aspects. First, we are interested to discover segments and not individuals. Second, we use data transformation methods to increase the search space. Third, we present some domain specific measures to assess the validity of the clusters. Finally, we propose visualisation methods to both investigate and compare cluster results to enable experts to compare and adapt their strategies.

In this paper, we present this architecture, that is based on the systemic search of dissimilar but relevant clusters. This approach consists of the generation of thousands of clusters by both transforming the input and using clustering methods with different parameters. These clusters are later evaluated to estimate how useful the clustering results are, and only the most coherent and the most different models are retained. Finally, the selected set of segments is provided to experts with structural information to facilitate the exploration and the comparison of the different segmentations.

The rest of this paper is organised as follows: The Customer Segmentation Architecture and the component that compose this architecture are described in the Section 2. Experimental results with discussion are provided in Section 3, and the conclusion of this paper and perspectives are given in Section 4.

II. Customer Segmentation Architecture

Our objective is to produce automatically diverse and meaningful customer segmentations. These segmentations will be used to help experts to discover specificities in consumer purchase behaviours.

Similarly to [10], we argue that producing multiple models may be adequate in order to offer experts alternative models. However we state that the produced models should respect two properties. First, all of the models output have to be of interest. Consequently, segmentation results should be assessed with at least one validity measure that will be later provided to experts. Second, all of the models have to be significantly different in terms of cluster composition or in terms of number of clusters. Moreover these differences should be highlighted to enable experts to select and to use the most adequate model with their strategy or to analyse the segment specificities to induce new knowledge. Although the number of different clusters is very high (the number of possible clustering is the Bell number), these constraints limit the segmentation results that will be kept and provided to experts.

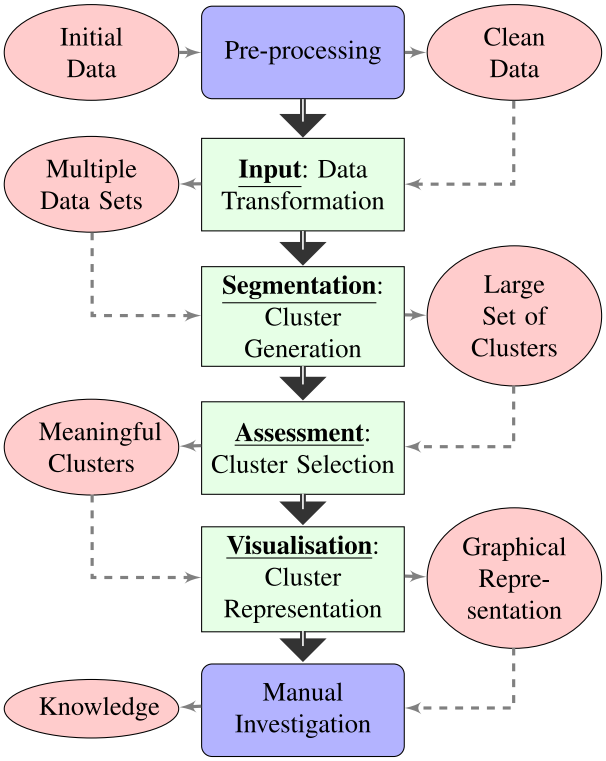


Figure 1. Architecture to discover consumer behaviours

Figure II presents the proposed architecture to create and select these diverse segmentations.

The automatic components are described in the following subsections. The manual steps and the parameters of the all the components are described in the next Section.

A. Input Component

The *input* component is in charge of the data transformation. This component is composed of several data transformation methods and may reduce or increase the number of features of the original data set. Moreover, all these transformations techniques may be combined to produce multi-transformed data.

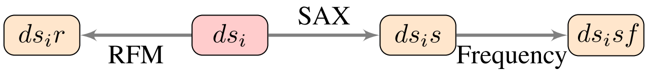


Figure 2. Data sets created from ds_i by the combination of transformation methods.

A first data set, ds_ir, is created to gather the Recency, the Frequency and the Monetary values of each customer. The Recency (R) describes the the last purchase recency given time t while the Frequency (F) and the Monetary (M) values represent respectively the number of purchases made and the amount of money spent. ds_ir consists of only the triplet ¡R, F, M¿, this means the 60 dimensions are reduced by a factor of 20. This data transformation is an example of the usage of implicit knowledge to transform data and extend the searched space.

A second set of data sets is created through the discretisation of the consumer data with the Symbolic Aggregate approXimation (SAX) method [14]. First, the time is discretised and data are separated into w periods. For each period, the value retained is the average value of the period under consideration. Second, an alphabet of size α is defined over the possible values, the amount in our case, and these values are replaced by the alphabet letters. As defined in SAX, the values are discretised such as the probability of each symbol is identical. The data set ds_is is composed of w dimensions, each of them being defined over an alphabet of size α.

Raw data provides information on simultaneous purchases. Such information are important as they describe how customers have reacted on special offers. It may also quantify how external events, such as Christmas, impact the brands sales. However, relations in the purchase frequencies or the similarities in the consumption rates may be of greater interest to segment the customers. Consequently, we create a last set of data sets that record the frequency transitions between the symbols identified by SAX. The frequency transition matrix holds the frequency of purchasing an amount A_t at time t and purchasing an amount A_{t+1} the following period. This method accept a parameter δ that represents the lag taken into account to compute the frequency transition. For example, let an alphabet α be of size 2 : $\alpha = \{0, 1\}$ and δ be 2, then the frequency transition table will be composed of eight values : $f_{00} \rightarrow 0, f_{00} \rightarrow 1, f_{01} \rightarrow 0, f_{01} \rightarrow 1, f_{10} \rightarrow 0, f_{10} \rightarrow 1, f_{11} \rightarrow 0, f_{11} \rightarrow 1$.

The number of features of this data set is $\alpha^{\delta+1}$ where α is the alphabet size and δ is the historic period taken into account.

Finally, each of these transformed data sets is sent to the following component.

B. Generation Component

The *generation* component is responsible of the creation of the clustering results. It is defined to accept several clustering methods with multiple parameters. When the clustering algorithm presents a stochastic behaviour, it is repeated r times.

We restrict the choice of the clustering algorithms to the algorithms that produce a hard partition. This limitation is due to the usage that will be made of this segmentation. Using a fuzzy clustering results may be very useful if it is combined with another learning technique. However, it is of a limited help if an expert has to explore and analyse the partition.

For each data set, multiple clustering are performed with various parameters. Similarly to the previous component, the clustering results may be produced iteratively or as a batch process.

C. Selection Component

The *selection* component's aim is to score the clustering results to keep the more relevant segmentations only.

We used three different approaches to quantify the quality of the clusters.

The first estimator, Q_1 is based on the coefficient of determination and measures the proportion of variability in the data set that is explained by the model and it is defined as :

$$Q_1 = R^2 = 1 - \frac{\sum(p_i - \mu_j)^2}{\sum(p_i - \bar{p})^2} \qquad (1)$$

where p represents the purchase information data and μ_j the cluster model associated to the customer i.

The number of clusters has an impact on Q_1 : with only one cluster, $Q_1 = 1$ while with n clusters and a population of n customers, the value of Q_1 is 0. Consequently, the comparison of clustering results with different number of clusters should not be performed directly. However, if required, methods such as the Gap-Statistic may discover the inflexion point and to select the most adequate size.

The second estimator, Q_2, is based on a classification task where the aim is to create segment with both homogeneous population and with identical RFM values at period $t + 1$ given the period t RFM value . The RFM Value is obtained from the R, F and M values of customers (RFM Value = R * F * M).

Then, the customers are ordered by their RFM values and spread into five quantiles of equal population.Usually RFM values of customers are consistent with the pyramid model that states that 20% of the customers account for 80% of the benefit. Finally customers received a label (very low, low, average, high, very high) provided the quantile they belong to, see Figure 3.

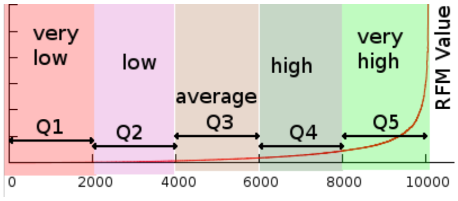


Figure 3. Labels of the five quantiles given the RFM Value

The clusters are then assessed by the dispersion over the labels among the clusters. We select the F-Measure, the harmonic mean of the precision and the recall to measure the homogeneity of the clusters. The Precision $P(l, c)$ is the proportion of the customers with the label l in the cluster c. The Recall $R(l, c)$ is the number of the customers with the label l in the cluster c over the total number of customers with label l.

Because the F-Measure combines Recall (that favours small clusters) and Precision (that favours large clusters), clustering results with different number of clusters can be compared.

The last clustering quality estimator, Q_3, indicates the accuracy of sales forecast by using the segmentation. The forecast is performed with the exponential smoothing. The sales at time $t+1$ depends on both the last real value and the last smoothed value. Given a real value v_t, a smoothed value s_t at time t and the smoothing parameter α, the forecast f_{t+1} value at time $t+1$ is given by :

$$f_{t+1} = \alpha v_t + (1-\alpha)s_t \tag{2}$$

The parameter α is selected for each segment by internal validation. The error is estimated with the Symmetric Mean Absolute Percentage Error (sMAPE) measure over the testing period (P weeks) :

$$sMAPE = \frac{1}{P}\sum_{t=1}^{P} 2\frac{|v_t - f_t|}{v_t + f_t} \tag{3}$$

The accuracy of the set of clusters C, given the measure Q_3, is then defined as :

$$Q_3 = 1 - sMAPE(C) \tag{4}$$

This result will be compared with the accuracy of the best prediction made on the whole population ($Q_3(n)$) or made by considering the customers individually ($Q_3(1)$). Solutions with a different number of cluster can be compared as Q_3 evaluates the clustering performance with an indirect measure.

Moreover, in addition to the actual quality of the segmentation and as recommended in [7], the segmentation accuracy over time is taken into account. To assess the segmentation consistency, we measure the segmentation quality when the model trained on ds_{train} is applied on the unknown data ds_{test}. Consequently, we divide the data sets into two parts. The training part, ds_{train} will be used to perform the segmentation, and the testing part will only be used to assess the clustering results.

For each of the defined estimator Q_i, we define a consistency measure C_{Q_i} :

$$C_{Q_i} = Q_i(train) \times Q_i(test) \tag{5}$$

We also compute a global coherency estimator, G, that takes into account the quality and the consistency measures :

$$G = C_{Q_1} \times C_{Q_2} \times C_{Q_3} \tag{6}$$

G consist of a pool of measure and similarly to [13], this architecture could be extended such as each measure receive a weigth that describes its contribution in the discovery of good segments.

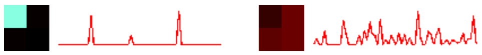
For all the clustering, the best results given each of the estimator Q_1, Q_2, Q_3, and G are selected and sent to the next component.

D. Visualisation Component

The *visualisation* component is in charge of the graphical representation of the clustering results. It has to provide information both on the clusters and on the estimated quality. Moreover, it has to provide tools to facilitate the comparison between different clustering results.

Different techniques are used to create the graphical representations. We present only the two most interesting, namely the Intelligent Icons representation and the formalism we proposed to highlight the differences between the clusters.

Intelligent Icons [15] are a technique that map the frequency transition between symbols into colours. Then given a frequency matrix, a squared icon can be derived to represent visually the matrix content. Recalling that we have discretised the data with SAX and calculated the frequency transition matrix (ds_isf), we can applied the same process to describe and visually represent the identified clusters.



(a) Customer 27 (b) Customer 94

Figure 4. Intelligent Icons applied on consumer purchase data (brand 1)

The Figure 4 demonstrates the efficiency of this representation for two different customers. The purchase and non-purchase events are represented by a square composed of four pixels. Although the information retained has been divided by 15, it still allow the comparison of two consumers.

Finally, to compare efficiently two different segmentations, the differences and the similarities in the population of the two segmentations must be highlighted. Consequently, the graphical representation should 1) shows the elements that appear together in the different segmentation and 2) similarly shows the elements that are separated in the two segmentations. The solution implemented use Intelligent Icons to represent these specific sub populations. Information on the size of the sub populations are also given together with a measure of similarity between the two population. An example of this graphical representation is given in Figure **??**.

III. Experimental Results

This section presents the methodology we used to perform the experimentations and the results obtained.

A. Methodology

Experiments were carried out on a data set obtained from the SLDS09 challenge [1]. This data set consists of the weekly

[1] Symposium Apprentissage et Science des Données 2009, www.ceremade.dauphine.fr/SLDS2009

purchase log of 10 000 customers over 62 weeks. Purchases were made for 3 brands in two different supermarkets (6 brands in total). The initial objective was to identify brand and/or supermarket specificities. The following investigation is performed only with the log of customer purchases : no additional information is known about the customers nor the brands.

Name	Value	Constraint
w	$w \in \{1, 2, 5, 10\}$	
α	$\alpha \in \{3, 4, 5, 6\}$	
δ	$\delta \in \{1, 2\}$	$\alpha^{\delta+1} < 60$

Table I
PARAMETERS USED TO TRANSFORM DATA

As shown in Figure II, the initial step, the preprocessing, is a manual phase. We removed the first and the last weeks because we noted that these weeks behave very differently from the others and we suspect a discrepancy in the way the data were collected. We transform the original data set in order to process each brand separately and obtain six independent data sets : $ds_1, ..., ds_6$. As explained in the last section, the data set were divided into two parts. The training set is composed of the first 50 weeks while the testing set holds the last 10 week purchase records.

The component *input* in charge of the data transformation is composed of different transformation methods that requires parameters. The SAX discretisation requires two parameters : w and α, respectively the length of a period and the alphabet size. To find the sequence transition frequency given the discretised values requires a parameter δ that describes the period length of the transitions. The table I presents the parameters used by the data transformation methods. After the diverse transformations, 32 data sets are available per brand.

To implement the component *segmentation*, we use the K-Means algorithm and set K to be in $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25\}$. The algorithm was repeated 10 times (r = 10). Altogether, for each data set we performed 120 segmentations, leading to a total of over three thousands clustering results per brand.

B. Results

Because of the lack of space, we only present and compare results with K = 5. Results for different number of clusters will be put online [2].

First, we present the R^2 results on the 6 brands. On average, the R^2 scores are very low, indicating that very few variance may be explained by the model. However, we can note two distinct behaviours in the population of the identified clusters : some individuals from different clusters have differences based on their consumption level or on the consumption periods. For example, segments of the brands 1, 3, 4 and 5 are clearly separated by the purchase volume. However, when considering the brand 2 and 6, we can see that clusters seems also to have temporal specificities. This indicates that clusters may regroup individuals that made simultanous purchases.

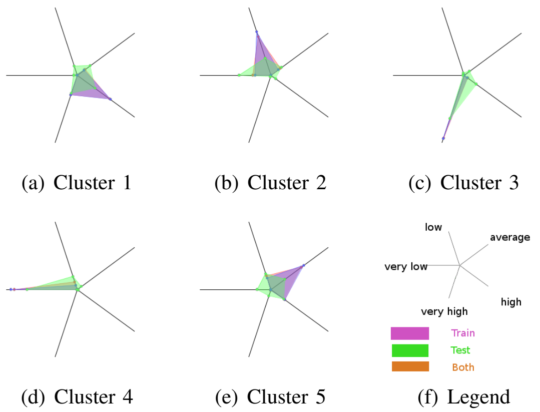
We can also notice that brands 2 and 6 (where segments are partially based on simulatenous purchases) obtain a better R^2 value with segmentation on $ds_i s$, while the others (where identified segments are separated by consumption rates) rely on the transition matrix data $ds_i sf$.

The results of the RFM Value classification, assessed by the F-Measure are given in the Table II. It is interesting to note that for the brand 1, 3, 4 and 5, the segmentation that separates the best the customers with their RFM values does not use the RFM data but the transition matrix.

Brand	1	2	3	4	5	6
Train	71.00	78.14	77.02	70.41	74.86	77.28
Test	40.96	48.56	43.41	39.94	41.56	48.47
Both	67.09	71.53	70.14	65.84	64.88	71.16
Data	$ds_i sf$	$ds_i r$	$ds_i sf$	$ds_i sf$	$ds_i sf$	$ds_i r$

Table II
BEST F-MEASURE ON LABELLED DATA WITH K = 5

Web Spider graphs are given for the best segmentation of size 5 on the brand 1 with respect to Q_2. These graphs show the internal class distribution and the evolution on both the training, the testing and the whole data set (from the first to the sixtieth week). Although clusters are not pure, the graphs show that each of the five clusters contains a certain type of customers. For example, the cluster 4 contains customers with a very low RFM value. We can note that segments with the customers of extreme RFM values (with very low or very high value) are more consistent.



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

(d) Cluster 4 (e) Cluster 5 (f) Legend

Figure 5. Best F-Measure Clustering (brand 0), with K = 5

When considering sale predictions, the difference with one

[2] Extra information are available on http://www.emining.fr/data/consumer-behaviour/

segment and n segments is not significant. This result is very disappointing as Q_3 may be a very sensible indicator of the actual segmentation performance. This result indicates that 1) exponential smoothing is not an adapted to perform the forecast on these data or 2) the selection of the smoothing parameter α is not performed properly, *i.e.* the interval validation should be take a longer historic.

When analysing the results we can see that the estimators Q_1 and Q_2 on the training data set are correlated with the testing data set while Q_3 results are not. Therefore, as such Q_3 is not consistent over time.

The predictive power of the segmentation should then be assessed through a more reliable forecasting method, such as methods based on neural networks.

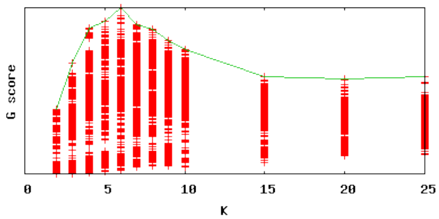


Figure 6. Global score for all the segmentations made on brand 1

Furthermore, the global score (Eq. 5) is the last method to perform the selection among the different clusters, with respect to all of the estimators defined. The Figure 6 shows the global score distribution for all the clustering results of the brand 1. On this figure, the best global segmentation is found with $K = 6$.

From this result, we can represent the segment behaviour by Intelligent Icons. Moreover experts can navigate through solutions by comparing intelligent icons together.

IV. Conclusion

In this paper we have presented a generic architecture to perform a customer segmentation on purchase log data by 1) transforming the data, 2) generating diverse models, 3) selecting the most adequate models given a set of evaluating functions and 4) creating a visual representation of the segmentations. This architecture has been used on a real-world data set and numerous diverse segmentation models have been produced, the best of them have been retained and graphically represented. In future work we intend to improve the sale forecasting and obtain more significant and more stable results, a forecasting method based on Neural Networks will be used instead of the method based on the exponential smoothing. Additionally, to increase the segmentation diversity, we plan to complement the experiments by using different clustering algorithms such as density-based methods. Finally, in this paper we have concentrated the investigations on intra bands segmentation. A further analysis could perform the segmentation based on customer inter-brand relations.

References

[1] Rui Xu and D. Wunsch. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678, 2005.

[2] M. McDonald. The role of marketing in creating customer value. In *Marketing from an Engineering Perspective (Digest No. 1996/172)*, pages 1/1–111, Nov 1996.

[3] Girish Punj and David W. Stewart. Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research*, 20(2):134–148, 1983.

[4] Vasilis Aggelis and Dimitris Christodoulakis. Customer clustering using rfm analysis. Stevens Point, Wisconsin, USA, 2005. WSEAS.

[5] W. Niyagas, A. Srivihok, and S. l Kitisin. Clustering e-banking customer using data mining and marketing segmentation. *ECTI*, 2(1), May 2006.

[6] C. Dennis, D. Marsland, T. Cockett, and V. Hlupic. Market segmentation and customer knowledge for shopping centers. In *ITI, 2003.*, June 2003.

[7] Roger J. Calantone and Alan G. Sawyer. The stability of benefit segments. *Journal of Marketing Research*, 15(3):395–404, 1978.

[8] Guozheng Zhang. Customer segmentation based on survival character. In *WiCom*, Sept. 2007.

[9] Jing Wu and Zheng Lin. Research on customer segmentation model by clustering. In *ICEC*, New York, NY, USA, 2005. ACM.

[10] Ching-Hsue Cheng and You-Shyang Chen. Classifying the segmentation of customer value via rfm model and rs theory. *Expert Syst. Appl.*, 36(3):4176–4184, 2009.

[11] Pei-Chann Chang, Chen-Hao Liu, and Chin-Yuan Fan. Data clustering and fuzzy neural network for sales forecasting: A case study in printed circuit board industry. *Knowledge-Based Systems*, 22(5):344–355, 2009.

[12] Tianyi Jiang and A. Tuzhilin. Improving personalization solutions through optimal segmentation of customer bases. In *ICDM*, Dec. 2006.

[13] Graham J. Williams. Evolutionary hot spots data mining - an architecture for exploring for interesting discoveries. In *PAKDD*, pages 184–193, 1999.

[14] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *DMKD*, New York, NY, USA, 2003. ACM.

[15] E. Keogh, Li Wei, Xiaopeng Xi, S. Lonardi, Jin Shieh, and S. Sirowy. Intelligent icons: Integrating lite-weight data mining and visualization into gui operating systems. In *ICDM*, Dec. 2006.