CUSTOMIZED GARMENT FASHION RECOMMENDATION SYSTEM USING DATA MINING TECHNIQUES

Shukla Sharma[123], Ludovic Koehl[12], Pascal Bruniaux[12], Xianyi Zeng[12]

[1]GEMTEX
[2]ENSAIT
[3]ECOLE CENTRALE DE LILLE
Lille, France

ABSTRACT

Many fashion firms have enabled their business model to give extremely personalized experiences to their customers by using advanced CAD tools like CLO 3D, Marvelous-Designer, Browzwear, Lectra and many more for designing the garment and build a 3D avatar for the customized garment as well as web-based services to be integrated with the web and mobile-based applications. Due to the integration of highly advanced technologies for designing and giving personalized experience has increased the customer's expectations. In this paper, we have presented our initial work to build a garment fashion recommendation system for customized garments, which can be used with mobile and web applications. The proposed system structure is designed on the user's biometric profile and historical data of product order. We have collected the user's historical data from a fashion company dealing with customized made-to-measure garments. Proposed architecture for recommendation system is based on different data mining techniques like clustering, classification and association mining.

KEYWORDS

Recommendation System, BIRCH, Adaptive Random Forest, Incremental learning, data mining, Association mining

1. INTRODUCTION

Digitalization in the fashion industry has automated each step by using CAD based tools as well as web-based platforms and made it easy to build the digital supply chain. Due to the digitalization fashion brands has generated a good amount of information to understand the consumers at various level and opened the door for the adoption of data-driven services based on machine learning-based data analysis. Recently -commerce platform Zalando has opted for data-driven marketing techniques based on artificial intelligence and for the accomplishment of building data-driven services has acquired AI start-ups to work on their business domain [1].

Big market players Amazon, Alibaba, Flipkart, Myntra have also invested to build a strong digital supply chain to grab more customers and fulfill their demand by understanding them with deep

Natarajan Meghanathan et al. (Eds) : CSEIT, CMLA, NeTCOM, CIoT, SPM, NCS, WiMoNe, Graph-hoc - 2019
pp. 373-385, 2019. © CS & IT-CSCP 2019 DOI: 10.5121/csit.2019.91329

analysis using Artificial Intelligence based services [1]. Deloitte analysis for fashion companies has shown that consumers are willing to bear the expensive amount for highly personalized garments and fashion accessories [2]. Highly personalized garments need a close connection between the consumer and designers and to build the recommendation system to fullfill the gap by considering the customized garment data available recommendation filtering techniques analyzed. The selection of the filtering technique depends on the domain of business and type of available data attributes [25]. Broadly RS can be divided into three categories first is content-based, second is collaborative and third is hybrid filtering techniques. Content-based (CB) filtering technique prepares the recommendation by matching the similarity with the content of the user's preferable items. Also, CB filtering method matches the content of the user's profile by considering attributes that show the interest of users. In the literature of fashion recommendation systems [3] it can be seen that the content-based analytic tools used to know the fashion brand's sales activity. Content-based filtering is not suitable where content for product and user profile is not well-formed. The content-based system fails to provide recommendations by considering product popularity. On the other hand, collaborative filtering (CF) builds the recommendation result by knowing the relationship between the users and items by considering item ratings. Collaborative filtering further can be divided into two types, first is item-based and second is user-based CF. User-based CF works on user rating given to the item and matching corresponding user profiles. The user-based approach becomes insufficient when data sparsity is seen. A small number of rated items between user profiles causes to generate unreliable and poor recommendations. Another issue is the user's profile information gets updated frequently and it requires the recompilation of user-based models. Also, an expectation to have ratings or item interaction data for every item leads CF-based systems to face item cold-start problems.

Another popular Item-based CF was introduced by Amazon to overcome the issues faced in user-based filtering techniques. The item-based filtering technique prepares prediction by calculating the similarity between items. The similarity between items remains stable as compare to user profile similarity and less requirement to recompile the item-based model. The item-based system builds recommendations for users by matching the similarity between the items and items purchased together with the large user population [4]. A collaborative filtering technique with visually explainable recommendations in the fashion industry has been proposed in the literature of recommendation systems.[5]. Item-based CF technique also inherits cold start problems when there is sufficient data to analyze the user's historical behavior for new item sets.

Hybrid filtering(HF) is a combination of various recommendation techniques to overcome the shortcomings of previously introduced Content-based and collaborative filtering. Applying a combination of machine learning algorithms or data mining techniques with a combined approach which is followed in content-based and collaborative filtering can improve the accuracy of the recommendation model. Because the issue of one algorithm can be solved by the other algorithm. In the literature of fashion recommendation systems, hybrid recommendation system is used to find the latest fashion trends [6]. We used a hybrid recommendation system using data mining techniques to analyze the customized shirt's data. Further paper is organized as follows. The next section 2 explains and highlights the existing fashion industrial report and recent issues faced by fashion e-commerce platforms regarding the design of garment. Section 3 we have shown data mining steps with a brief introduction of incremental clustering algorithms. Further sub-section under section 3 shows details about the BIRCH clustering, Adaptive Random Forest classification and association mining algorithms and their results. The last section concludes the activities completed in this research paper.

2. PROBLEM STATEMENT FOR CUSTOMIZED GARMENT RECOMMENDATION SYSTEM

A significant increase and adoption of the CAD tools have given a very closed personalized experience to consumers. Fashion companies are using the latest technologies like virtual AR VR platforms, Mirrors.

A virtual dressing room by GAP provides customers virtual garments over 3D avatars of the user's body morphology. CAD tools helped to design garment designs suited to different body type and 3D visualization made it easier to see the garment's ease over the body. Famous CAD tools like Lectra, ClO3D are used by designers to create the garment pattern and then evaluate the fitting on 3D avatar. Despite having so much detailed information and virtual dressing rooms still, there is a gap between the designers and users' understanding related to fashion attributes. The famous sales drop issue in Zalando has shown that returns are not only because of fittings but also related to garment design and fabric choice for a garment. Stacia Carr Zalando's Director of Engineering mentioned that design and fabric also can have an impact on the returns. Zalando raised alarm to one popular denim brand that faced sales drop due to a small change in the design [7]. In this research paper, to build a fashion recommendation system using each and every attribute of the customized shirt like collar cuff, pocket, fabric, color and biometric profile of the user. Hybrid filtering with data mining techniques is used to build the system. Because the available dataset doesn't fulfill the requirements of content-based and collaborative filtering techniques. Both the filtering techniques needs product description and rating data and rating data collection for a customized garment is not feasible due to the multiple numbers of style attributes of the garment.

Asking users for rating all visible attributes is the complex and not user-friendly approach, for example, shirt attribute collar can have many different types of collar styles and the same applies to other style attributes like shirt button type, pocket type, cuff type, back yoke type. A web platform for customized garments works more closely with complex body measurement techniques because the individual body has its own specificity[8]. Through our research work, we aim to explore different scenario for a customized garment from designer's and consumer's point of view and we contribute as follows:

- Build a close connection between users and designers by analyzing consumer's selection choices for different styles and attributes available on the system.
- GFRS is proposed to give a real-time recommendation by handling incremental data.

Following steps for building hybrid recommendation systems are as follows:

- Recommendation model composition starts from data partitioning by taking the vector of biometric parameters specific to customized garments.
- Data partitioning divides data into small segments each segment contains similar biometric profiles of users.
- Each segment is further classified corresponding to the fitting type.
- Frequent itemset patterns mined by applying association learning to the different variants of a shirt. Frequent itemsets are the combination of attributes selected together for a shirt.

3. METHODOLOGY

Data mining is the process of finding the hidden patterns from large data sets and generate useful insights for business. The development of a recommendation system is specific to. Every domain has different needs of recommendation systems like a proposed system in [9] has focused on fashion company which sells products through online web platforms as well as with offline showrooms. Therefore building the recommendation system with a conventional model used in other e-commerce websites and social networks can not fulfill the requirement of the different types of fashion business models.

3.1. Data Mining

Data mining is not a new term and drills down data to get useful insights for business have a long history. The extraction of the hidden pattern accelerates the pace of decision-makers. The exponential use of latest technologies and the proliferation of data in the fashion industry has raised the need for using big data analysis techniques to get valuable insights for fashion companies [10].

Commonly used clustering algorithms are categorized as partitioning, hierarchical, grid-based, and model-based algorithm [11] Partitioning methods require the value of k for algorithm by a user and it relocates the instances from one cluster to another. K-mean is most commonly used in the algorithm. k-mean partitions data with their respective centroids. The centroid is calculated by taking the mean of all instances in the cluster. K-medoids is another partitioning algorithm also known as PAM (Partition around medoids) Hierarchical based clustering algorithms are further categorized into two types:

- Agglomerative hierarchical clustering follows a bottom-up approach. Initially, each element is considered as a cluster of its own. Every iteration works to combine most similar instances into a bigger cluster node. The iteration process works until the desired cluster structure is formed.
- Divisive hierarchical clustering - follows a top-down approach. All instances are considered as a single cluster at first step and then moving forward with each iteration most heterogeneous cluster of instances divided into sub-clusters and it continues until every object is assigned to their cluster.

Grid-based clustering algorithm CLIQUE is a gird based cluster to work with high dimensional data. It helps to find the cluster in the subspace of high dimensional data and it does not require to select the subspace which might have a cluster. Because of automatic subspace clustering, CLIQUE has been recommended to use in data mining applications[12].

Model-based clustering algorithms use a different model for each cluster and tries to find the best fitting for that model. Statistical learning and neural-based learning are two types of model-based learning. Statistical learning includes COBWEB, GMM, and neural-based learning includes SOM, ART [13]. In our research work, we have clustered data using BIRCH (balanced iterative reducing and clustering using hierarchies) algorithm.

BIRCH algorithm is an unsupervised hierarchical algorithm for clustering. BIRCH is suitable for handling large data sets and incrementally and dynamically handle new metric data points [14]. The BIRCH clustering algorithm completes the process in two steps. The first step is building CF Tree and loads data into a cluster feature tree(CF Tree).CF Tree is a compressed form of data. BIRCH algorithm becomes highly efficient by using summary statistics for minimizing large data sets. CF Tree is built with CFs and each CF is composed of three summary statistics [14]:

- The count represents the number of data values in the cluster.
- The linear sum is the sum of individual coordinates and helps to measure the location of the cluster.
- Squared Sum is the sum of squared coordinates and helps to measure the spread of the cluster.

The second step is clustering the sub-clusters. After the creation of the CF Tree existing clustering algorithm on CF Tree Leaf nodes(sub-clusters) is applied to combine sub-cluster into clusters.

3.2. Dataset Description and Recommendation Model Building Steps

We have collected data from European fashion companies that are dedicated to creating a customized garment. We have considered customized shirts dataset for building an initial model. Detailed description related to data can be seen in Table 1, Table 2, Table 3 GFRS model is designed by combining three steps:

Table 1. Customized shirt data

.

Month count	Order Count	User Count	Remake Request Count
10	5291	4712	493

Table 2. Customized shirt design attributes

Fabric	Fit	Collar	Placket	Cuff	Pocket	CollarWhite	CuffWhite
331	8	44	14	32	10	2	2

Table 3. user biometric profile

Height	Weight	CollarSize	Age

Data Clustering: Data clustering is used as a pre-processing step, Which becomes useful at later steps of data analysis. Also makes easier to build initial overview on the dataset by applying statistical analysis and machine learning algorithms [15] BIRCH algorithm is used to create the separate groups which are based on user biometric profile for the shirt. Three-dimensional input vector [height, weight, collar size] has passed to get the data in homogeneous segments for further analysis. In this step we have created segments of similar objects for our recommendation model. The reason for doing segmentation on the basis of biometric profile is useful to deal with the extremely different and similar biometric profile of users for example extremely long height

and weight user profiles won't be suitable for group of users who are having extremely short height and weight. Following steps used for the completion of clustering:

- We have used the Scikit machine learning library to implement the BIRCH algorithm for online clustering.
- BIRCH model is trained in 5 iterations where each iteration is containing 1000 approx records to analyze the incremental behavior of the model.
- Silhouette Coefficient, Davies-Bouldin score metrics are used to evaluate the clusters in each iteration.

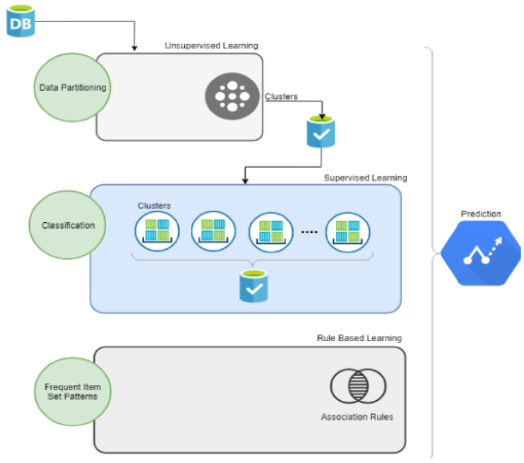

Figure 1. Initial architecture of our Garment Fashion Recommendation System.

Cluster evaluation Silhouette Coefficient metric is used to know how well clusters are formed. It is calculated using mean intra-cluster distance and mean nearest cluster distance [16]. Its value ranges from -1 to 1. Coefficient value 1 is considered as the best value and indicates the good clusters formed. If the coefficient value is going towards 0 it means that formed clusters are overlapping and if the value is -1 then this negative value indicates that samples have been assigned to the wrong cluster or clusters are not well-formed [17]. Davies-Bouldin score's minimum value is 0 which is considered to be good and it is calculated as the average similarity score of each cluster with its most similar cluster.

A similarity measure is calculated as the ratio between within-cluster distances and between cluster distances [17]. Table 4 shows the Silhouette Coefficient metrics increasing as new data is used to train the model cluster formation is reducing the overlap between the clusters. The silhouette score moving towards 1 is considered good and 0 indicates that clusters are overlapping. A negative score means the wrong assignment of samples to the cluster.

Table 4. Cluster evaluation

Iteration	silhouette_coefficient	davies_bouldin_score
0	0.327517934551341	0.98885472069335
1	0.322808536550035	0.80856055643328
2	0.294895935082953	1.01848641365009
3	0.311195152561083	1.02506770111566
4	0.307172835023094	1.04154161015915

Davies Bouldin score is the average similarity score of each cluster with its most similar cluster. decreasing and going towards 0 shows clusters are not similar and less dispersed.

Data Classification: We have used supervised learning in this step using Adaptive random forest classifier to predict best fitting type class for input vector [X, Y] where X containing three biometric parameters [height, weight, collarsize] Y contains fitting type as a label. The adaptive random forest (ARF) classifier is suitable for online learning or to follow an incremental approach. Traditional classifier needs to be retrained if new data comes by losing previous information or appending new data with an older dataset and retrain the model to use all information. Bagging, Random forests are an example of ensemble classifiers. Bagging predicts by aggregating the multiple version of predictors [18]. Random forest is another popular and widely used ensemble classifier that uses a forest of trees and they vote for the popular class [19]. Increase the performance by combining the prediction of all models. ARF is an updated version of random Forest Following characteristics from experiments done in a research paper[20]:

- ARF obtains good classification performance on real-world data.
- A large number of feature handling using a small number of trees.
- ARF can train its base trees in parallel without affecting its classification performance. This is an implementation concern, but it is useful for investigating scalability.
- ARF might not be able to improve on data sets where all features are necessary to build a reasonable model.

Table 5. Classification report for Cluster 1

class_name	f1score	precision	recall	support
3	0,078431373	1	0,040816327	49
4	0	0	0	1
7	0,717557252	0,598726115	0,895238095	210
8	0,32	0,5	0,235294118	102
9	0	0	0	2
accuracy	0,587912088	0,587912088	0,587912088	0,587912088
macro avg	0,223197725	0,419745223	0,234269708	364
weighted avg	0,514203737	0,620144187	0,587912088	364

Table 6. Classification report for Cluster 2

class_name	f1score	precision	recall	support
1	0	0	0	3
3	0,795275591	0,677852349	0,961904762	105
4	0	0	0	3
5	0	0	0	1
7	0,383561644	0,736842105	0,259259259	54
8	0	0	0	2
accuracy	0,68452381	0,68452381	0,68452381	0,68452381
macro avg	0,196472872	0,235782409	0,203527337	168
weighted avg	0,620334915	0,660499823	0,68452381	168

Table 7. Classification report for Cluster 3

class_name	f1score	precision	recall	support
1	0,425531915	0,5	0,37037037	27
3	0	0	0	12
4	0	0	0	3
5	0,4	1	0,25	12
7	0,627345845	0,527027027	0,774834437	151
8	0,401869159	0,494252874	0,338582677	127
accuracy	0,521084337	0,521084337	0,521084337	0,521084337
macro avg	0,309124486	0,420213317	0,288964581	332
weighted avg	0,488120384	0,505575892	0,521084337	332

Table 8. Classification report for Cluster 4

class_name	f1score	precision	recall	support
1	0,476190476	0,5	0,454545455	11
3	0,171428571	0,666666667	0,098360656	61
4	0,133333333	0,5	0,076923077	13
5	0	0	0	5
7	0,653846154	0,488505747	0,988372093	86
8	0,1	1	0,052631579	19
9	0	0	0	1
accuracy	0,5	0,5	0,5	0,5
macro avg	0,219256934	0,450738916	0,238690408	196
weighted avg	0,38550684	0,579990617	0,5	196

Classification Evaluation: In the second step, we trained 4 classifiers based on each group identified in the clustering step. The classification report generated to see the quality of the model's prediction. Classification model evaluation values corresponding to each metric can be seen in Table 5, Table 6, Table 7, Table 8. F1 score, precision, recall, support metrics are used to know the model's prediction over different classes.. Precision shows the classifier's ability to not label positive instance as negative or percentage of prediction was correct. Precision value is calculated for each class by dividing true positives by the sum of true and false positives.

$$tp/(tp + f\,p) \quad (1)$$

A recall is the percentage of positive cases in the classifier model and for each class, the ratio is calculated by dividing true positive by sum of true positives and false negatives.

$$tp/(tp + f\,n) \quad (2)$$

Support shows the number of actual occurrences of the class In the specified dataset.

Classification report for some classes f1score value is 0 which gives an indication true positive + false positive == 0 or true positive + false negative == 0 .Also classes showing 0 value for precision means that there were no true positive case found for that class.

Prequential Evaluation: Interleaved test then train is used to test the incremental Adaptive Random Forest Classifier's accuracy with different pre-trained instances Figure 1 shows mean performance accuracy of classifier increased gradually as the number of instances increased to pre - trained the model.

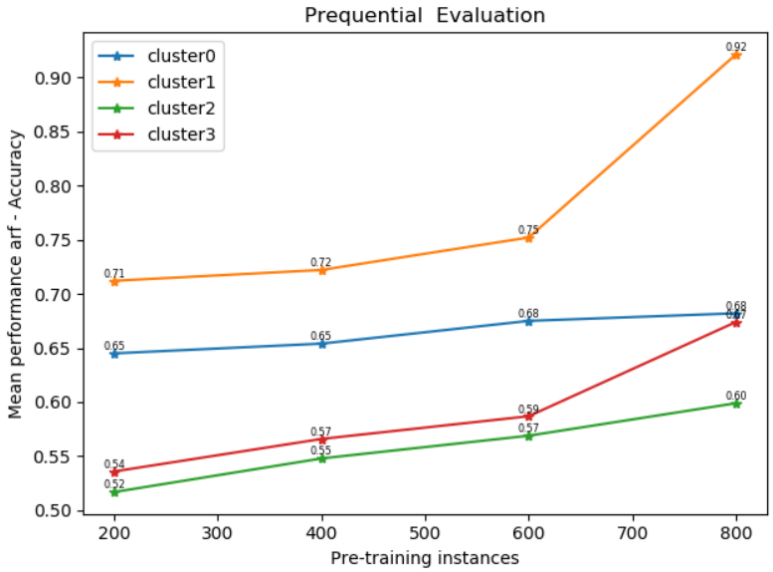

Figure 1. ARF performance accuracy calculation using prequential evaluation

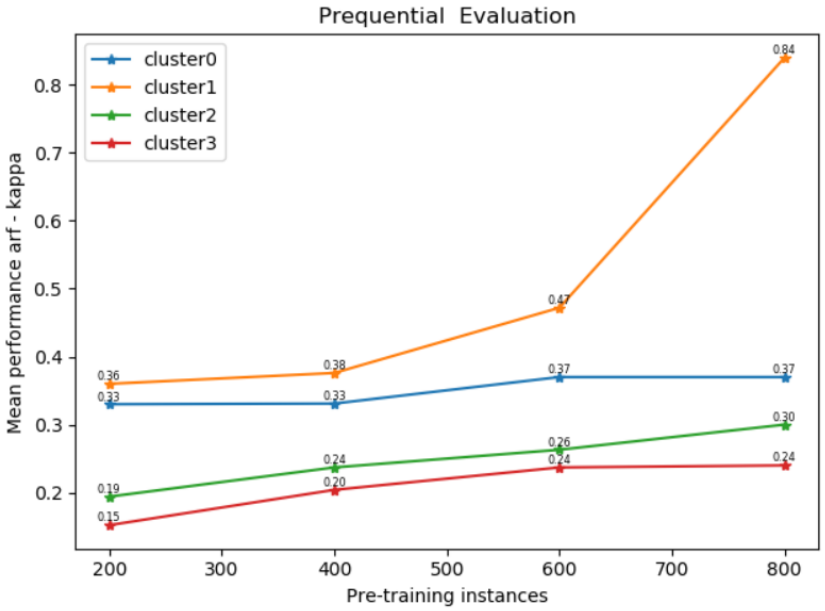

Figure 2. ARF kappa score calculation using prequential evaluation

Figure 2. Shows the kappa score for ARF classifier in a moderate range. Kappa score ranges between -1 to 1 and used to measure inter-annotator agreement, negative score and score close to zero shows chance agreement and 1 means complete agreement.

Association Mining: Association mining has used in the third step to getting the frequent itemset and association rules for the shirt attributes. In the literature of the recommendation system association mining is considered for building a web-based personalized recommendation system where explicit feedback is missing or can not be collected from users [21]. Also, association mining has been used in previous research work for improving the web-based application's performance by getting the best set of rules using the Apriori algorithm [22]. Association mining process needs frequent itemsets to build rules and Apriori is one of the famous widely used algorithms. It was first introduced in 1996 [23] for finding all frequent itemsets using the Apriori candidate generation procedure and next is to get the support of frequent itemsets is counted. Apriori's candidate generation process becomes costly when there are long patterns. On the other hand, FP tree algorithm is an efficient algorithm to find frequent itemsets for long patterns because of its compact tree data structure which helps to avoid repeated data scans [24]. FP Growth is faster than the Apriori algorithm.

FP Growth with minimum support value 0.1 has trained with all set of attributes Fabric, Collar, Cuff, Placket, Pocket, CollarWhite, CuffWhite to get the frequent pattern sets. Getting a number of frequent pattern set depends on the value of minimum support as value goes down a number of frequent itemsets increases. Support, confidence, and lift metrics have been used to filter interesting rules. Association rules are represented as A → C, where A is antecedent itemset and C is consequent itemset. Support is the combined metric for antecedent itemset and consequent itemset and gives a percentage of transactions that contain both antecedent and consequent. The confidence value for rule is to know the truthness of consequent occurrence corresponding to antecedent in transaction database. Lift value greater than 1 indicates that antecedent and consequents are dependent and if the value is 1 two sets are independent of each other and can't be considered as potential rules.

Table 6. Filtered association rules

antecedents	consequents	support	confidence	lift
{'no_collarwhite',Classic Point_collar})	{'no_remake'})	0.156374502	0.945783133	1.02544953
{'Classic Point_collar',no_remake})	{'no_collarwhite'})	0.156374502	0.951515152	1.016299162
{'no_collarwhite',Double inc. Cufflinks_cuff})	{'no_remake'})	0.230079681	0.950617284	1.030690878
{'Double inc. Cufflinks_cuff',no_remake})	{'no_collarwhite'})	0.230079681	0.954545455	1.019535783
{'n0_collarwhite',Hai Cutaway_collar})	{'no_remake'})	0.120517928	0.930769231	1.009170959
{'no_remake',Hai Cutaway_collar})	{'no_collarwhite'})	0.120517928	0.952755906	1.017624393
{'Round Single_cuff',no_remake})	{'no_collarwhite'})	0.310756972	0.939759036	1.00374263
{'Italian Semi-Spread_collar',no_remake})	{'no_collarwhite'})	0.162350598	0.942196532	1.006346083

Machine learning library: scikit-multiflow open-source framework for machine learning is used to implement BIRCH, Adaptive random forest algorithm. Association mining is implemented using MLxTEND machine learning library. In the future we proposed recommendation model will be implemented as web service and will be tested with online web and mobile-based platform.

4. CONCLUSIONS

We have presented initial steps for building Garment fashion recommendation system by considering biometric profiles to segregate the data. Also customer's style preference corresponding to different fitting style has known by generating association rules. Proposed recommendation system has combined unsupervised, supervised and association learning to build the GFRS. GFRS system is developed to track user preferences corresponding to their selection behaviour on design elements corresponding to body measurement profile. Because customized garment online platforms are very specific to body type. BIRCH,Adaptive random forest classifier is used to make scaled system and to provide nearly real time prediction. In our future research work we will work with other customized garments to make this system more generalized. Recommendation service will be developed and will be tested with the online platform which connects different fashion brands(Project partners) and validates their business cases.

ACKNOWLEDGMENTS

We thank the European Union for providing an opportunity to contribute FBD BMODEL H2020 project

REFERENCES

[1] Achim Berg, Imran Amed, Anita Balchandani, Johanna Andersson, Saskia Hedrich, and Robb Young. (2019) Fashion industry trends to watch in 2019 |McKinsey. [Online]. Available: https://www.mckinsey.com/industries/retail/ourinsights/ten-trends-for-the-fashion-industry-to-watch-in-2019

[2] N. Wixcey. (2015) Made to order: The rise of mass personalisation | deloitte switzerland | consumer business. [Online]. Available:https://www2.deloitte.com/ch/en/pages/consumer-business/articles/madeto-order-the-rise-of-mass-personalisation.html

[3] B. Touchette, Morgan Schanski, and Seung-Eun Lee, "Apparel brandsâĂŹ use of facebook: an exploratory content analysis of branded entertainment | emerald insight," vol. Vol. 19 No. 2, pp. 107-119. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/JFMM-04-2013-0051/full/html

[4] G. Linden, B. Smith, and J. York, "Amazon. com recommendations: Item-to-item collaborative filtering," IEEE Internet computing, no. 1, pp. 76–80, 2003.

[5] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, and H. Zha, "Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation," in Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR'19. New York, NY, USA: ACM, 2019, pp.765–774. [Online]. Available: http://doi.acm.org/10.1145/3331184.3331254

[6] Ã. Cardoso, F. Daolio, and S. Vargas, “Product characterisation towards ,: Learning attributes from unstructured data to recommend fashion products,” in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD ’18. ACM Press, pp. 80–89. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3219819.3219888

[7] Stacia Carr. Online clothing retailers hunt for better fit to cut costly returns - reuters. [Online]. Available: https://www.reuters.com/article/us-onlineapparel-returns-focus/online clothing-retailers-hunt-for-better-fit-to-cut-costly-returns-idUSKCN1OK1E2

[8] A. Kim, “Method of ordering a custom-made suit online,” Feb. 28 2019, uS Patent App. 15/687,690.

[9] H. Hwangbo, Y. S. Kim, and K. J. Cha, “Recommendation system development for fashion retail e-commerce,” Electronic Commerce Research and Applications, vol. 28, pp. 94–101, 2018.

[10] I. Amed, Johanna, ersson, A. Berg, M. Drageset, S. Hedrich, and S. Kappelmark. The state of fashion 2018: Renewed optimism for the fashion industry | McKinsey. [Online]. Available: https://www.mckinsey.com/industries/retail/ourinsights/ renewed-optimism-for-the-fashion-industry

[11] A. Bhattacharya, N. Chowdhury, and R. K De, “Comparative analysis of clustering and biclustering algorithms for grouping of genes: co-function and co-regulation,” Current Bioinformatics, vol. 7, no. 1, pp. 63–76, 2012.

[12] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, “Automatic subspace clustering of high dimensional data for data mining applications,” SIGMOD Rec., vol. 27, no. 2, pp. 94–105, Jun. 1998. [Online]. Available: http: //doi.acm.org/10.1145/276305.276314

[13] D. Xu and Y. Tian, “A comprehensive survey of clustering algorithms,” vol. 2, no. 2, pp. 165–193. [Online]. Available: https://doi.org/10.1007/s40745-015-0040-1

[14] T. Zhang, R. Ramakrishnan, and M. Livny, “Birch: An efficient data clustering method for very large databases,” SIGMOD Rec., vol. 25, no. 2, pp. 103–114, Jun. 1996. [Online]. Available: http://doi.acm.org/10.1145/235968.233324

[15] B. Karmakar and I. Mukhopadhayay, “An efficient partition-repetition approach in clustering of big data,” in Big Data Analytics. Springer, 2016, pp. 75–93.

[16] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” vol. 20, pp. 53 – 65. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0377042787901257

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[18] L. Breiman, “Bagging predictors,” vol. 24, no. 2, pp. 123–140. [Online]. Available: https://doi.org/10.1007/BF00058655 [19] ——, “Random forests,” vol. 45, no. 1, pp. 5–32. [Online]. Available: https: //doi.org/10.1023/A:1010933404324

[20] H. M. Gomes, A. Bifet, J. Read, J. P. Barddal, F. Enembreck, B. Pfharinger, G. Holmes, and T. Abdessalem, “Adaptive random forests for evolving data stream classification,” vol. 106, no. 9, pp. 1469–1495. [Online]. Available: https://doi.org/10.1007/s10994-017-5642-8

[21] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, "Effective personalization based on association rule discovery from web usage data," in Proceedings of the 3rd International Workshop on Web Information and Data Management, ser. WIDM '01. ACM, pp. 9–15, event-place: Atlanta, Georgia, USA. [Online]. Available: http://doi.acm.org/10.1145/502932.502935

[22] S. Bayati, A. F. Nejad, S. Kharazmi, and A. Bahreininejad, "Using association rule mining to improve semantic web services composition performance," in Control and Communication 2009 2nd International Conference on Computer, pp. 1–5.

[23] R. Agrawal and J. C. Shafer, "Parallel mining of association rules," IEEE Transactions on Knowledge & Data Engineering, no. 6, pp. 962–969, 1996.

[24] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, ser. SIGMOD '00. ACM, pp. 1–12, event-place: Dallas, Texas, USA. [Online]. Available: http://doi.acm.org/10.1145/342009.335372

[25] M. D. Buhmann et al., 'Recommender Systems', in Encyclopedia of Machine Learning, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2011, pp. 829–838.

AUTHORS

Shukla Sharma, Ph.D. Candidate, GEMTEX Laboratory, France