

A Survey on Wine Dataset Classification

Bhavya AG

Department of Computer Science and Engineering
Cambridge Institute of Technology KR Puram
Bengaluru, Karnataka, India
Email: bhavyaga59@gmail.com

Abstract – As of late, the majority of the ventures advancing their items dependent on the quality accreditation they got on the items. The customary method for evaluating the item quality is tedious, anyway with the create of AI strategies the procedures has gotten more productive and expended less time than previously. In this paper we have investigated, a portion of the AI strategies to survey the nature of wine dependent on the characteristics of wine that relies upon quality. We have utilized white wine and red wine quality dataset for this exploration work. The wine informational collection has been being used in examine for quite a while and still it stays as the benchmark informational collection. Nature of wines is hard to characterize as there are numerous components that impact the apparent quality. This paper introduces a basic survey of research slants on Wine quality and usercentric similitude quantifies also. A tale client driven comparability measure in item bunching is proposed to assess the mainstream Wine informational index named Red Wine dataset. The exploratory outcomes got in this work can give preferable proposals to item purchasers over the current frameworks. The proposed methodology is equipped to amass the Red wine dataset into requested gatherings of favored wine variations and can pass judgment on the wine quality dependent on these client inclination gatherings.

Keywords– wine informational, AI strategies

I. INTRODUCTION

The intrinsic characteristics (visual, taste, smell), environmental characteristics (climate, region, site) and management practices (viticulture practice), as well as hysicochemical ingredients (acid, pH, etc.) are the factors f interest in assessing the quality of Wine. Data mining echniques in predicting wine quality are in progress, with ome promising results in the domain. Physicochemical and sensory tests are crucial in Wine certification. It is the routine practice in physicochemical laboratory tests, to characterize wine by determination of density, alcohol or pH values, but sensory tests rely mainly on human experts.

Wine classification is a difficult task as taste is the least understood of the human senses. The relationship between the physicochemical and sensory analysis are complex to understand. In the food industry, in addition to the food quality research, machine learning techniques have also been applied in classification of wine quality. Machine learning methods provide the way to build models from data of known class labels to predict the quality of a wine. In old days Wine was considered as a luxury item. Today, it is popular and enjoyed by a wide variety of people. Professional wine reviews offer insights on wines available n large quantity in each year. A systematic way is needed to utilize those large number reviews to benefit wine

Consumers, distributors, and makers. No two persons judge he wine alike even they taste the wine simultaneously while being able to share and detect all the same attributes. Experience helps a lot and hinders the taster. So assessing the quality of wine depending only on the taster's experience and sensing is a big process.

In recent years there is a modest increase in the wine consumption as it has been found that wine consumption has a ositive correlation to the heart rate variability [1]. With the ncrease in the consumption wine industries are looking for lternatives to produce good quality wine at less cost. Different wines have different purposes. Although most of the chemicals are same for different type of wine based on the chemical tests, the quantity of each chemicals have different level of concentration for different type of wine. These days it is really important to classify different wine for quality assurance [2]. In the past due to lack of technological resources it become difficult for most of the industries to classify the wines based on the chemical analysis as it takes lot of time and also need more money. These days with the advent of the machine learning techniques it is possible to classify the wines as well as it is possible to figure out the importance of each chemical analysis parameters in the wine and which one to ignore for reduction of cost.

II. RELATED WORK

In the past few attempts have been made to use different machine learning approaches and feature selection techniques

to the wine dataset. Er and Atasoy proposed a method to classify the quality of wines using three different classifier such as support vector machines, Random forest and k-nearest neighborhood. They have used principal component analysis for feature selection and they found good result using Random forest algorithm [3].

Chen et al proposed an approach that will predict the grade of wine using the human savory reviews. They have used hierarchical clustering approach and association rule algorithm to process the reviews and predict the wine grade and they found an accuracy of 85.25% while predicting the grade [4].

Appalasamy et al proposed a method to predict wine quality based on physicochemical test data. They have pointed out that classification approach helps to improve the quality of wine during the production [5].

Beltrán et al proposed an approach to classify the wine based on aroma chromatograms and they have used PCA for dimensionality reduction and wavelet transform for feature extraction and classifiers such as neural network, linear discriminant analysis and support vector machine and found that support vector machine with wavelet transforms perform better than other classifiers [6].

Thakkar et al., used analytical hierarchy process (ahp) to rank the attributes and then used different machine learning classifiers such as support vector machine and random forest and they found accuracy of 70.33% using random forest and 66.54% using SVM [7].

Reddy and Govindarajulu used a user centric clustering approach to recommend the product. They have used red wine data set for the survey purpose. They have allocated relative voting to the attributes based on the literature review. Then they assigned weight to the attributes using Gaussian Distribution Process. They judged the quality based on the user preference group [8].

The above past work motivated us to try different feature selection algorithm as well as different classifiers to compare the performance metrics. This paper proposed GA based feature selection and SA based feature selection and used different classifiers such as PART, RPART, Bagging, C5.0, random forest, svm, lda, naïve bayes etc.

A good number of research papers have been published on wine quality that is mostly based on the empirical studies in the wine industry. Most of the research was endorsed to assess wine quality using physicochemical data is based on small sample sizes. In [18] pattern recognition approaches that include clustering, principle component analysis, nearest neighbors, etc. were applied

to classify wines from Galicia (northwestern Spain) among several different brands.

The dataset used consists of 42 white wines. Principle component analysis (PCA) for wine classification according to the geographical region was reported in [21].

The authors used the data set that contains 33 Greek wines with physicochemical variables. The details of a 2-stage classification done (principle component and clustering) from 24 industrial fermentations of a particular type of wine were given in [9].

This study tried to detect undesirable fermentation behavior. In [14] authors proposed an improved KNN method with weights, which considerably improves the performance of KNN method. The authors employed a kind of preprocessing on train data. They introduced a new value named Validity to train samples which cause to more information about the situation of training data samples in the attribute space. This new value takes into accounts the value of stability and robustness of any train samples regarding with its neighbors. KNN with applied weights employs validity as the multiplication factor yields to more robust classification rather than simple KNN method, efficiently. The Wine dataset is the result of a chemical analysis of wines grown in the same region in Italy that derived from three different cultivars. The analysis determined the quantities of 13 attributes found in each of the three types of wines. This data set has been in use with many others for comparing various classifiers. In the context of classification, this is a well-posed problem with wellbehaved class structures. Data mining techniques to classify the quality of wines using a larger physicochemical data set were used in more recent works. Cortez and his colleagues [20] built models using support vector machine, multiple regression, and neural networks. A dataset with a large number of records is considered (vinho Verde samples from the Minho region of Portugal). A computational procedure was developed that performs simultaneous variable and model selection. Support vector machine achieved desirable results, "outperforming the multiple regression and neural network methods". This model is vital in supporting the oenologist wine tasting evaluations and to improve wine production. The results of this research are relevant to the wine science domain, helping in the understanding of physicochemical characterization and the things that affect the final quality.

In[15] authors enforced unsupervised neural network (NN) based on Adaptive Resonance Theory (ART1) as an alternative to statistical classifier so as to discriminate among the 178 samples of wine possessing 13 numbers of attributes. The dimensionality of the feature variables was reduced to 5 by principal component analysis (PCA). Out of 13, the first 2 numbers of principal components

captured over 55.4 % of the variance of the wine dataset. Nonhierarchical K-means clustering algorithm was used to choose the classes available among the samples of wine. Appalasaamy et al [10] applied two classification algorithms, decision tree and Naive Bayes and compared results with the recent work results.

In [11] authors proposed a brand new data science area named Wine informatics. In order to automatically retrieve wines' flavors and characteristics from reviews, which are stored in the human language format, authors proposed a novel "Computational Wine Wheel" to extract keywords. Two completely different public-available datasets are produced based on the new technique in their paper. The hierarchical clustering algorithm is applied to the primary dataset and got purposeful clustering results. Association rules mining is performed on the second dataset to predict whether or not a wine is scored higher than 90 points or not supported on the wine savory reviews. Fivefold crossvalidation experiments were executed based on different parameters and results with a range of 73% to 82% accuracy were generated. This new domain will bring huge benefits to fields as diverse as computer science, statistics, business, and agriculture.

In [13] authors proposed a data analysis approach to classify wine into different quality categories. A data set of white wines of 4898 records was used in the analysis. As the data set was imbalanced with about 93% of the observations are from one category with respect to the occurrence of events in it, Synthetic Minority Over-Sampling Technique (SMOTE) was applied to oversample the minority class. A balanced data was considered to model a classifier that categorizes a wine into three categories.

These categories include high quality, normal quality, and poor quality. Three classification techniques used in this work include decision tree, adaptive boosting and random forest. Among the techniques, random forest produced to produce the desired results with the minimal error. Based a Wine dataset of 4898 instances the authors attempt to build models that classify different wines into quality categories. With this model, the test data is tested. The quality variable is assessed by many factors The authors concluded that the analysis would give a clearer idea to winemakers as to which variables influence the quality the most and what steps could be an attempt to attain more desirable outcomes.

In [24] authors used Analytical Hierarchy process (AHP) classification algorithm. This algorithm provides the way to recommend wine on the basis of the components of the wine. Wine selection on the basis of its attributes is a different approach. The Machine Learning Techniques used here helped in finding the component accuracy of wine attributes. The Analytical Hierarchy Process (AHP) is used for arranging and examining complicated

problems by mathematical calculations. AHP defines multiple attribute issue to advise a particular commodity to an individual. The process of AHP is mainly used to calculate weights. The inputs for AHP are relative preferences and attributes. The authors have taken red wine dataset and weights were allotted to them based on AHP. The obtained results after analysis of the data were used for recommending a wine to individuals. Users need reasonably good recommendations in finding products according to their likes and dislikes.

In [22] authors compare data-centric evaluation with user-centric evaluation and obtain remarkable results in favor of usercentric approach.

In [12] authors used a new clustering algorithm based on the mutual vote, which adjusts itself automatically to the given dataset, needs minimum overhead in terms of parameters, and also able to detect clusters with different densities in the same dataset. Currently, many Voting/Rating machine searing based automated recommender software systems are developing continuously by many companies for voting based selection of products. A customer would be interested in purchasing products that are similar to the products that he/she liked earlier.

In [23] authors proposed an algorithm that combines user-based approach, item-based and Bhattacharyya approach. The main advantage of this hybrid approach is its capability to find more reliable items for recommendation. Collaborative filtering technology works by creating a database of preferences for products by their customers. Collaborative technology is becoming popular in the latest research areas such as E-business, Banking, Space, and Share Market and so on.

In [19] authors proposed an improved collaborative filtering algorithm that combines k-means algorithm with CHARM algorithm. This hybrid approach improved the prediction quality of recommendation system.

In [17] authors tried to improve the learning speed by splitting the cluster tree into sub-clusters and by using exploration and exploitation phases and aggregates as well. From user-centric sensor data, Friendbook discovers lifestyles of users and measures the similarity of lifestyles between users. It recommends friends to users if their lifestyles have high similarity.

In [24] authors model the daily life of users as life documents. Using these documents lifestyles are extracted by using the Latent Dirichlet Allocation algorithm.

III. CONCLUSION

In this paper, As of late, the majority of the ventures advancing their items dependent on the quality

accreditation they got on the items. The customary method for evaluating the item quality is tedious, anyway with the create of AI strategies the procedures has gotten more productive and expended less time than previously. In this paper we have investigated, a portion of the AI strategies to survey the nature of wine dependent on the characteristics of wine that relies upon quality. We have utilized white wine and red wine quality dataset for this exploration work.

The wine informational collection has been being used in examine for quite a while and still it stays as the benchmark informational collection. Nature of wines is hard to characterize as there are numerous components that impact the apparent quality. This paper introduces a basic survey of research slants on Wine quality and usercentric similitude quantifies also. A tale client driven comparability measure in item bunching is proposed to assess the mainstream Wine informational index named Red Wine dataset. The exploratory outcomes got in this work can give preferable proposals to item purchasers over the current frameworks. The proposed methodology is equipped to amass the Red wine dataset into requested gatherings of favored wine variations and can pass judgment on the wine quality dependent on these client inclination gatherings.

IV. FUTURE WORK

In future we can try other performance measures and other machine learning techniques for better comparison on results. This analysis will help the industries to predict the quality of the different type of wines based on certain attributes and also it will helpful for them to make good product in the future.

REFERENCES

- [1]. I.Janszky, M.Ericson, M.Blom, A. Georgiades, J.O.Magnusson, H.Alinagizadeh, and S.Ahnve, "Wine drinking is associated with increased heart rate variability in women with coronary heart disease," *Heart*, 91(3), pp.314-318,2005.
- [2]. V. Preedy, and M. L. R. Mendez, "Wine Applications with Electronic Noses," in *Electronic Noses and Tongues in Food Science*, Cambridge, MA, USA: Academic Press, 2016, pp. 137-151.
- [3]. Y.Er, and A.Atasoy, "The Classification of White Wine and Red Wine According to Their Physicochemical Qualities,"*International Journal of Intelligent Systems and Applications in Engineering*,4,pp.23-26,2016.
- [4]. B. Chen, C. Rhodes, A. Crawford, and L. Hambuchen, "Wineinformatics: applying data mining on wine sensory reviews processed by the computational wine wheel," *IEEE International Conference on Data Mining Workshop*, pp. 142 149, Dec. 2014.
- [5]. P.Appalasamy, A.Mustapha, N.D.Rizal, F.Johari, and A.F.Mansor, "Classification-based Data Mining Approach for Quality Control in Wine Production," *Journal of Applied Sciences*, 12(6), pp.598-601,2012
- [6]. N. H. Beltran, M. A. Duarte- MERMOUND, V. A. S. Vicencio, S. A. Salah, and M. A. Bustos, "Chilean wine classification using volatile organic compounds data obtained with a fast GC analyzer," *Instrum. Measurement, IEEE Trans.*, 57: 2421-2436, 2008.
- [7]. K.Thakkar,J.Shah,R.Prabhakar,A.Narayan,A.Joshi, "AHP and machine Learning techniques for wine recommendataions" , *Irnternational Journal of computer science and information technologies*, 7(5), pp. 2349-2352, 2016.
- [8]. Reddy, Y. S., & Govindarajulu, P. (2017). An Efficient User Centric Clustering Approach for Product Recommendation Based on Majority Voting: A Case Study on Wine Data Set. *IJCSNS*, 17(10), 103.
- [9]. [9] Alejandra Urtubia, J Ricardo Perez-Correa, Alvaro Soto, and Philippe Pszczolkowski. "Using data mining techniques to predict industrial wine problem fermentations". *Food Control*, 18(12):1512–1517, 2007.
- [10]. Appalasamy P, A Mustapha, N. D. Rizal, F Johari, and A. F. Mansor. "Classification-based data mining approach for quality control in wine production". *Journal of Applied Sciences*, 12(6):598–601, 2012.
- [11]. Bernard Chen, Christopher Rhodes and Aaron Crawford, "Wineinformatics: Applying Data Mining on Wine Sensory Reviews Processed by the Computational Wine Wheel",*IEEE International Conference on Data Mining Workshop*.
- [12]. Charif Haydar, Anne Boyer "A New Statistical Density Clustering Algorithm based on Mutual Vote and Subjective
- [13]. Logic Applied to Recommender Systems" *UMAP'17*, July 9-12, 2017, Bratislava, Slovakia.
- [14]. Gongzhu Hu, Tan Xi, Faraz Mohammed, "Classification of Wine Quality with Imbalanced Data", 2016 IEEE.
- [15]. Hamid Parvin, Hoseinali Alizadeh, Behrouz Minati," A Modification on K-Nearest Neighbor Classifier" *GJCST*, Vol.10 Issue 14 (Ver.1.0) November 2010.
- [16]. Kavuri N. C. and Madhusree Kundu. "ART1 Network: Application in Wine Classification", *International Journal of Chemical Engineering and Applications*, Vol. 2, No. 3, June 2011.
- [17]. Kunal Thakkar et al, (IJCSIT),"AHP and Machine Learning Techniques for Wine Recommendation" *International Journal of Computer Science and Information Technologies*, Vol. 7 (5), 2016, 2349-2352.
- [18]. Linqi Song, Cem Tekin, Mihaela van der Schaar "Clustering Based Online Learning in Recommender Systems: A Bandit Approach", *SongTekin ICASSP2014*.

- [19]. Maria J Latorre, Carmen Garcia-Jares, Bernard Medina, and Carlos Herrero, "Pattern recognition analysis applied to classification of wines from Galicia (northwestern Spain) with the certified brand of origin", *Journal of Agricultural and Food Chemistry*, 42(7):1451–1455, 1994.
- [20]. Paritosh Nagarnaik, Prof. A. Thomas, "Survey on Recommendation System Methods", IEEE SPONSORED 2ND INTERNATIONAL CONFERENCE ON ELECTRONICS AND COMMUNICATION SYSTEM (ICECS 2015).
- [21]. Paulo Cortez, Antonio Cerdeira, Fernando Almeida, Telmo Matos, and Jos'e Reis. "Modeling wine preferences by data mining from physicochemical properties". *Decision Support Systems*, 47(4):547–553, 2009.
- [22]. S Kallithraka, IS Arvanitoyannis, P Kefalas, An El-Zajouli, E Soufleros, and E Psarra. "Instrumental and sensory analysis of Greek wines; implementation of principal component analysis (PCA) for classification according to geographical origin". *Food Chemistry*, 73(4):501–514, 2001.
- [23]. Soude Fazeli, Hendrik Drachslar, Marlies Bitter-Rijpkema, Francis Brouns, Wim van der Vegt, and Peter B. Sloep, "User-centric Evaluation of Recommender Systems in Social Learning Platforms: Accuracy is Just the Tip of the Iceberg", *IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES*.
- [24]. Umutoni Nadine, Huiying Cao, Jiangzhou Deng, "Competitive Recommendation Algorithm for E-commerce" 2016, 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD).
- [25]. Zhibo Wang, Jilong Liao, Qing Cao, Hairong Qi, Zhi Wang, "Friendbook: A Semantic-based Friend Recommendation System for Social Networks", *IEEE TRANSACTIONS ON MOBILE COMPUTING* 2015.