Wineinformatics: Applying Data Mining on Wine Sensory Reviews Processed by the Computational Wine Wheel

Bernard Chen  
University of Central Arkansas  
201 Donaghey Ave.  
Conway AR 72035 USA  
bchen@uca.edu

Christopher Rhodes  
University of Central Arkansas  
201 Donaghey Ave.  
Conway AR 72035 USA  
CTR06001@UCA.EDU

Aaron Crawford  
University of Central Arkansas  
201 Donaghey Ave.  
Conway AR 72035 USA  
AaronC@uca.edu

Lorri Hambuchen  
Pulaski Technical College  
3000 W Scenic Dr.  
North Little Rock AR 72118 USA  
lorrihambuchen@mac.com

ABSTRACT
As the world becomes more digital, data Science is the successful study that incorporates varying techniques and theories from distinct fields. Among all fields, the domain knowledge might be the most important since all data science researchers need to start with the domain problem, and end with useful information within the domain. Identifying new application domain is always considered as fundamental research in the area.

Wine was considered as a luxury in old days; however, it is popular and enjoyed by a wide variety of people today. Professional wine reviews provide insights on tens of thousands wines available each year. However, currently, there is no systematic way to utilize those large number reviews to benefit wine makers, distributors and consumers. This project proposes a brand new data science area named Wineinformatics. In order to automatically retrieve wines' flavors and characteristics from reviews, which are stored in the human language format, we propose a novel “Computational Wine Wheel” to extract key words. Two different public-available datasets are produced based on our new method in this paper. Hierarchical clustering algorithm is applied on the first dataset and retrieved meaningful clustering results. Association rules algorithm is performed on the second dataset to predict whether a wine is scored above 90 point or not based on the wine savory reviews. 5-fold cross validation experiments are executed based on different parameters and results with a range of 73%-82% accuracy are generated. This new domain will bring huge benefits to fields as diverse as computer science, statistics, business and agriculture.

Keywords
Wine Informatics, Data Science, Data Mining, Clustering, Association Rules

1. INTRODUCTION
Data Science is the study that incorporates varying techniques and theories from distinct fields, such as Data Mining, Scientific Methods, Math and Statistics, Visualization, natural language processing, and the Domain Knowledge, to discover useful information from domain-related data. Among all fields, the domain knowledge might be the most important since all data science researchers need to start with the domain problem, and end with useful information within the domain.
be processed by computers **automatically** through the computational wine wheel proposed in this paper. Clustering and association rule algorithms are also implemented and discussed in this paper. We cluster wines together based solely on the words used in professional wine reviews. An association rules algorithm is applied to find the relationship between the wine characteristics and its grade within a 100-points scale. Many other data mining algorithms and related researches can be easily applied in this new and exciting area. The rest of the paper is organized as follows: Section 2 describes the knowledge about wine and our input data; Section 3 provides the process of the data preprocessing; Section 4 demonstrates the application of the clustering and its results; Section 5 shows the usage of association rules to find meaningful rules and predict the quality of the wine; Finally, we include the conclusion and future works.

2. **WINE SENSORY DATA**
Wine sensory analysis involves tasting a wine and being able to accurately describe every component that makes it up. Not only does this include flavors and aromas, but characteristics such as acidity, tannin, weight, finish, and structure. Within each of those categories, there are multitudes of possible attributes or forms that each can take. What makes the wine tasting process so special is the ability for two people to simultaneously view the same wine differently while being able to share and detect all the same attributes. It is an art where experience can both help and hinder the taster, as even the most experienced taster finds new combinations or preparation techniques that challenge the wine making process that has been established for thousands of years.

2.1 **Wine Spectator**
We chose Wine Spectator as our primary data source because of their strong on-line wine review search database and consistent wine reviews. These reviews are mostly comprised of specific tasting notes and observations while avoiding superfluous anecdotes and non-related information. They review more than 15,000 wines per year and all tastings are conducted in private, under controlled conditions. Wines are always tasted blind, which means bottles are bagged and coded. Reviewers are told only the general type of wine and vintage. Price is also not taken into account. Their reviews are straight and to the point. Wine Spectator provides an effective tool for users to easily search their consistent and precise reviews.

For each reviewed wine, a rating within a 100-points scale is given to reflect how highly their reviewers regard each wine relative to other wines in its category and potential quality. The score summarizes a wine’s overall quality, while the testing note describe the wine’s style and character. The overall rating reflects the following information recommended by Wine Spectator about the wine [14]:

- **95-100** Classic: a great wine
- **90-94** Outstanding: a wine of superior character and style
- **85-89** Very good: a wine with special qualities
- **80-84** Good: a solid, well-made wine
- **75-79** Mediocre: a drinkable wine that may have minor flaws
- **50-74** Not recommended

Below is an example of one of the wine reviews from our dataset.

**Kosta Browne Pinot Noir Sonoma Coast 2009** 95pts
Ripe and deeply flavored, concentrated and well-structured, this full-bodied red offers a complex mix of black cherry, wild berry and raspberry fruit that’s pure and persistent, ending with a pebbly note and firm tannins.

2.2 **Wine Aroma Wheel**
In the review example listed above, the attributes are neatly stated without much confusion to what constitutes a proper wine tasting note. For example, just reading the whole review, we would pull out the following attributes: RIPE, DEEPLY FLAVORED, CONCENTRATED, WELL-STRUCTURED, RED, BLACK CHERRY, WILD BERRY, RASPBERRY, PURE, PERSISTENT, PEBBLY and FIRM TANNINS. We do this for all one hundred wines and store the complete list in a SQL database for easy querying. However, we needed to make sure the attributes we were pulling were accurate, so we based our initial observations off a popular Wine Aroma Wheel, made by retired professor and sensory chemist, Ann C. Noble. The Wine Aroma Wheel is an excellent tool as it is a multi-level diagram designed to allow people unfamiliar with some flavors and aromas to try and detect them while wine tasting. What it does for us though is gives us a base set of categories, subcategories, and specific wine tasting notes we can use while examining the wine reviews. Figure 1 is a sample Wine Aroma Wheel.

![Figure 1. Ann C. Noble’s Wine Aroma Wheel](image)

2.3 **Computational Wine Wheel**
While the Wine Aroma Wheel is a great start, we found it lacking in the sense that it does not capture non-flavor notes which always appear in the wine reviews, such as tannins, acidity, body, structure, or finish. We also wanted to capture whatever observations were made that did not fit into any of the above categories. For example, if wine was considered BEAUTIFUL or SMOOTH. These descriptions, while not actual tasting notes, still add a subtle amount of character to a wine that we wished to capture.

In order to develop our computational wine wheel that have the capability to accurately capture all keywords appear in wine reviews, we have to start with a small but representative list of wines. Fortunately, in the end of each year, Wine Spectator selects the Top 100 wines based on quality (represented by score), value (reflected by release price), availability (based on the number of cases either made or imported into the United States) and an “X-factor”, which is excitement. The minimum score to be considered
as a candidate for the Top 100 of the year is 90/100. In this paper, the first dataset is compiled from the list of “Top 100 Wines of 2011” [16] as we started this research in mid-2012. The goal is to capture all attributes demonstrated by representative wines of the year.

After carefully analyzing all one hundred wine reviews and adding all necessary categories and subcategories, we came out with a total of 547 distinct attributes. When looking at our finished list, we noticed many cases where groups of attributes were really just permutations of the same thing. Because of this, we introduce what we call Normalized Attributes to our dataset. An example would be the following three attributes: FRESHLY-CUT APPLE, RIPE APPLE, and APPLE. None of these three attributes offer any significant value on their own. So for our purposes, we normalize all three of these examples to the attribute, APPLE. This group normalization needs to be carefully examined though, as we would consider something like GREEN APPLE to be unique enough to stand out on its own. Luckily, we have a domain expert, (one of our author), in our team to help verify these attributes. She is a wine instructor at a Culinary School (name hidden from reviewers) where she helps professionals receive industrial wine certifications. After going through our complete list, we were able to cut our total attributes from 547, down to 376. The table 1 shows the final categories and subcategories we used in our dataset. For each combination, it also shows the distinct count of attributes versus normalized attributes. Due to the space limitation, we do not show all 376 attributes offer any significant value on their own. So for our purposes, we normalize all three of these examples to the attribute, APPLE. None of these three attributes were really just permeations of the same thing. Because of this, finished list, we noticed many cases where groups of attributes with a total of 547 distinct attributes. When looking at our dataset, which includes 100 wines, the average number of attributes per wine is 12.85 and the standard deviation is 2.73. The minimum attributes of a wine is 8 and the maximum is 23.

2.4 An Example of Using the Computational Wine Wheel

Table 2 Simplified computational wine wheel

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>SUBCATEGORY</th>
<th>ORIGINAL</th>
<th>NORMALIZED</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRUITY</td>
<td>TREE FRUIT</td>
<td>BERRY</td>
<td>BERRY</td>
</tr>
<tr>
<td>FRUITY</td>
<td>TREE FRUIT</td>
<td>WILD BERRY</td>
<td>WILD BERRY</td>
</tr>
<tr>
<td>FRUITY</td>
<td>TREE FRUIT</td>
<td>RASPBERRY</td>
<td>RASPBERRY</td>
</tr>
<tr>
<td>OVERALL</td>
<td>TANNINS</td>
<td>FIRM TANNINS</td>
<td>TANNINS_HIGH</td>
</tr>
<tr>
<td>OVERALL</td>
<td>TANNINS</td>
<td>CHWEY TANNINS</td>
<td>TANNINS_HIGH</td>
</tr>
<tr>
<td>OVERALL</td>
<td>TANNINS</td>
<td>LIGHT TANNINS</td>
<td>TANNINS_LOW</td>
</tr>
</tbody>
</table>

As shown in Table 1, most specific fruits and flavors were not affected a greatly by the normalization process. It was the opinionated observations in the FLAVOR/DESCRIPTORS subcategory that resulted in the most change. This was because we needed to make sure attributes such as DELICIOUS, DELICIOUSLY, and DELICIOUSNESS were all normalized into DELICIOUS instead of treating them differently. In our first dataset, which includes 100 wines, the average number of attributes per wine is 12.85 and the standard deviation is 2.73. The minimum attributes of a wine is 8 and the maximum is 23.

In order to clarify the usage of the computational wine wheel, we provide an example in this subsection. Table 2 gives a simplified computational wine wheel, which contains only 6 attributes. To compare with the computational wine wheel we proposed in this paper, the simplified version has only 6 original attributes and 5 normalized attributes (3 original and 3 normalized attributes in FRUITY category and TREE FRUIT subcategory; 3 original and 2 normalized attributes in OVERALL category and TANNINS subcategory). Here is the process of how we apply the simplified computational wine wheel on the following Kosta Brown Pinot Noir’s wine review:

**Kosta Browne Pinot Noir Sonoma Coast 2009** 95pts
Ripe and deeply flavored, concentrated and well-structured, this full-bodied red offers a complex mix of black cherry, wild berry and raspberry fruit that's pure and persistent, ending with a pebbly note and firm tannins.

Use the words in the ORIGINAL column, which is the 3rd column in table 2, to scan the review starting with the longest number of combination word. Since the longest number of combination word in the example is 2, we start with WILD BERRY, followed by FIRM TANNINS, CHWEY TANNINS and LIGHT TANNINS.

Every time we had a hit, the wine will have a positive attribute in the corresponding NORMALIZED attribute and remove the word from the review. Therefore, after the scan of two combination word, the Kosta Brown Ponit Noir will have the following attributes:

<table>
<thead>
<tr>
<th>BERRY</th>
<th>RASPBERRY</th>
<th>WILD BERRY</th>
<th>TANNINS HIGH</th>
<th>TANNINS LOW</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

And then we scan the review again with single word ORIGINAL attribute. The update attributes are:

<table>
<thead>
<tr>
<th>BERRY</th>
<th>RASPBERRY</th>
<th>WILD BERRY</th>
<th>TANNINS HIGH</th>
<th>TANNINS LOW</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Please note that the BERRY attribute is still negative since we delete the word “WILD BERRY” from the review during the previous scan. The readers may notice that many important attributes in the Kosta Brown Ponit Noir example are NOT included, such as BLACK CHERRY, RIPE, DEEPLY FLOWERED... etc. It is because the computational wine wheel is the simplified version. The more ORIGINAL and NORMALIZED words included in the computational wine wheel, the more attributes can be picked up from the wine reviews to produce more accurate results.

3. Wine Clustering
3.1 Distance Measurement
After the first dataset was generated, which contains 100 wines and 376 wine related attributes, we planned to apply clustering algorithms on it to find similar wines in 2011’s top 100 list. Needless to say, how to calculate the distance would be the first priority problem to solve. Since the attribute values are binary in nature (a wine either has an attribute or it does not), regular distance measurement such as Euclidian distance or Manhattan distance will not fit. We decided to use Jaccard’s Coefficient when comparing two wines:

$$Jaccard’s \ Coefficient = \frac{P}{P+Q+R}$$

$$Jaccard’s \ Distance = \frac{Q+R}{P+Q+R}$$

Where

- P = Number of variables positive for both objects
- Q = Number of variables positive in Q, but not R
- R = Number of variables positive in R, but not Q

A value of 1 in the coefficient is completely similar and a value of 0 is completely dissimilar. Table 3 provides an example for our distance calculation. This example includes three wines and four distinct attributes. The Jaccard’s coefficient between wine 1 and 2 equals to 1/6 (since P=1, Q=3+2, R=0); The Jaccard’s coefficient between wine 1 and 3 equals to 3/6 (since P=0, Q=3+2, R=0); The Jaccard’s coefficient between wine 2 and 3 equals to 2/6 (since P=0, Q=3+2, R=0).

Table 3. Example of wine distance calculation by Jaccard’s Coefficient

<table>
<thead>
<tr>
<th></th>
<th>BLUEBERRY</th>
<th>CHERRY</th>
<th>CHEWY TANNINS</th>
<th>BEAUTY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wine 1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Wine 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Wine 3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

However, in this example, our wine expert asserted that the similarity value in (wine 1, wine 2) and (wine 1, wine 3) should not be the same since the wine both with “cherry” flavor should be more similar than the wine both with “beauty”. This point is exaggerated since the wines we included in our first dataset are from top 100 list. Therefore, we introduce a weight system, which allows us to classify certain groups of attributes as being more important when making a binary attribute comparison. Table 4 shows the descriptions of each weight. In the weighted scheme, the formula for Jaccard’s coefficient stays the same, but with different P, Q and R definitions:

Table 4. Weighted Mechanism

<table>
<thead>
<tr>
<th>Weight</th>
<th>Attribute does not appear in Wine Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Least Important Attributes – non-flavor descriptions and non-flavor wine characteristics (pure/beauty/wonderful/etc...)</td>
</tr>
<tr>
<td>2</td>
<td>Semi-Important Attributes – non-flavor wine characteristics (tannins/acidity/body/etc...)</td>
</tr>
<tr>
<td>3</td>
<td>Important Attributes – food wine characteristics (specific fruit, woods, flavors, etc...)</td>
</tr>
</tbody>
</table>

Table 5 provides an example with weighted attributes. In the new example with the weighted system, the Jaccard’s coefficient between wine 1 and 2 equals to 1/6 (since P=1, Q=3+2, R=0); The Jaccard’s coefficient between wine 1 and 3 equals to 3/6 (since P=3, Q=2+1, R=0), which is much more meaningful than the previous example; The Jaccard’s coefficient between wine 2 and 3 equals to 2/6 (since P=0, Q=3+2, R=0).

Table 5. Example of wine distance calculation with weighted attributes by Jaccard’s Coefficient

<table>
<thead>
<tr>
<th>WEIGHT</th>
<th>BLUEBERRY</th>
<th>CHERRY</th>
<th>CHEWY TANNINS</th>
<th>BEAUTY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wine 1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Wine 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Wine 3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

3.2 Clustering Results
Hierarchical clustering creates a hierarchical decomposition of the input-datasets and represents it as a dendrogram. There are two major categories of the hierarchical clustering: agglomerative approach (bottom-up) and divisive approach (top-down) [20]. After Eisen et al [17] successfully applied hierarchical clustering into microarray gene expression data in graphic form; it has become the most popular clustering method in the area [18]. In this paper, we try to reach the same achievement by using the agglomerative hierarchical clustering (bottom-up) method with visualization as our analysis tool. All 100 wines with 376 attributes are fed into Hierarchical Clustering Explorer (HCE) [19]. We used Single-Link Hierarchical Clustering with weighted Jaccard’s Similarity Coefficient as the similarity measurement. Figure 2 gives the whole result of hierarchical clustering.
Once the hierarchical clustering is performed, where to “cut” the dendrogram to form the clusters is always an active research area. HCE is able to show the tree and from top to bottom, a horizontal cut line can be made to show clusters under 0% to 100% similarity, respectively. It is important to note that this 0% to 100% range is actually relative to the similarity measure of the two closest initial observations. For example, if the two closest observations are 80% similar, every other similarity measure will be less than this, so the similarity range in the dendrogram is 0% to 100% in relation to the “80%” cluster. Figure 3 shows the example of 11 clusters using HCE by having a horizontal cut at 0.600. This cut represents all clusters where every member has a 60% or greater similarity in relation to our most similar cluster, which has a Jaccard’s Similarity Coefficient of 0.444768 (44.7368%).

By looking into the example clusters, we show some interesting results. Reviews of the wines in cluster#5 are given below as an example:

**Argiano Toscana Non Confunditur 2009** 92pts
An appealing graphite aroma adds depth to the black currant, violet and cedar aromas and flavors in this svelte red, which is smooth, with finely woven tannins. The bright acidity drives the spicy finish. Cabernet Sauvignon, Merlot, Syrah and Sangiovese.

**Moccagatta Barbaresco Bric Balin 2007** 94pts
Super fresh and focused, delivering pure flavors of blackberry, black currant, violet and mineral, with a hint of tar. It's all backed by refined tannins and bright acidity, providing structure that should allow this to develop over the next 20 years.
Common attributes are colored in blue to show the similarities of the wine. Table 6 provides the common attributes for each cluster. There are tremendous ways to use the clustering results, for example: (1) this information is beneficial to the customers who know what specific wine features they would enjoy most. For example, a customer who likes wines with Plum, Mineral and long finish characteristics could be recommended the wines from cluster #1. (2) Also, the wines in a cluster share similar wine flavor attributes; however, they do not share the same price tag. The price of the wines in cluster #1 is ranging from $30 (both Januik Cabernet Sauvignon Columbia Valley 2008 and Tablas Creek Cotes de Tablas Paso Robles 2009) to $125 (Domaine Serene Pinot Noir Dundee Hills Grace Vineyard 2008). Customers can enjoy similar wines with lower prices based on the clustering results. (3) This clustering information is likewise beneficial to the wine sellers to make a typical recommendation such as “customers who like this wine will also enjoy this” when the customers purchase or highly rate a wine in a cluster. (4) Some websites, such as Lot 18 [21], let users pay a small fee to do “Palate Calibration” by sending them several varieties of wine samples and ask for their feedback. Based on the collected customers-feedback, wine experts could know the wine flavors that the customers would appreciate most. In this case, wine sellers can use table 6 to find the best match for personalized selection.

Table 6. Table of common attributes (share more than 50% among the cluster) exhibited in each cluster

<table>
<thead>
<tr>
<th>Cluster #</th>
<th>Number of wines</th>
<th>Common attributes and the number of wines share this attribute in the cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>Plum(10), Mineral(7), Long Finish(7), Tannins_Medium(6)</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Floral(2), Blackberry(2), Berry(2), Mineral(2), Firm(2)</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>Blackberry(6), Long_finish(5), Spice(4)</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>Spice(3), Raspberry(3), Tannins_medium(2), Black Cherry(2), Mineral(2), Smooth(2), Harmony(2), Rich(2), Long Finish(2)</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>Tannins_Medium(2), Acidity_High(2), Violet(2), Black Currant(2)</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>Ripe(2), Tannins_High(2), Mineral(2), Complex(2), Well-Structured(2), Raspberry(2)</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>Toasty Wood(2), Spice(2), Black Licorice(2), Full-Bodied(2), Pure(2), Finesse(2), Mineral(2)</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>Pepper(2), Spice(2), Complex(2), Full-Bodied(2), Sage(2)</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>Dense(2), Herbs(2), Mineral(2), Red(2), Smoke(2)</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>Spice(2), Fig(2), Finesse(2), Rich(2), Delicacy(2), Melon(2), Layers(2)</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>Peach(2), Mineral(2), Mango(2), Tangerine(2), Smooth(2)</td>
</tr>
</tbody>
</table>

Wines made from the same region or grape type may share some similar properties, thus, they may be clustered together. This assumption is also reflected in our clustering results since many clusters share common properties other than wine attributes. First of all, all clusters are either all red wines or all white wines. Secondly, many clusters have the same wine country and/or region: all wines in Cluster #2, 6, 7, 8, 10 are from California USA; All wines in Cluster #5 are from Italy; All wines in Cluster #9 are from Castilla y Leon, Spain. Thirdly, some clusters also capture the same grape types: All wines in Cluster #8 are made by the blend of Grenache, Petite Syrah, and Syrah; All wines in Cluster #10 are made by the grape type Chardonnay. All wines in Cluster #11 are made by the grape type Sauvignon Blanc.

4. Wine Classification

The purpose of this section is trying to correctly predict whether or not a wine is scored more than 90/100, which means this wine is either a classic or outstanding wine, by merely using wine review information. It might be easy for a human expert to read a review and determine if a wine is scored above 95; however, 90 point border is a gray area. For example:

CLOS DU MONT-OLIVET Chateauneuf-du-Pape White90pts
A fresh, crunchy style, with bouncy honeysuckle, honeyed and green fig flavors followed by a lively, minerally finish.

DE WETSCHOF Chardonnay Robertson Limestone Hill 89pts
Very fresh style, with bouncy pear skin, green fig and green apple fruit flavors coursing along, laced with a flash of verbena.

Both wines have nice reviews, but one is above 90 points and the other is below 90 points. Among all classification algorithms, we prefer “white box” algorithms (decision tree, association rules) over “black box” (SVM) ones. The transparent decision process contrasted by a model would reveal important information for wine makers’ references. We try to use association rules algorithm to handle this challenging problem in this section.

4.1 Association Rules in Wine Informatics

Association rules were initially made popular by market-basket analysis [22]. The algorithm was developed in order to discover the connection between items in a purchase for large transactional databases. While made popular initially for marketing strategies, the algorithm can be useful for finding many relational insights in data. The association rules algorithm generates the rules in a form of A=>B. The rule A=>B holds both support s, which is the probability that a transaction contains A∪B, and confidence c, which is the conditional probability that a transaction having A also contains B. The formula is given below:

\[
\text{Support}(A \Rightarrow B) = \frac{|A \cup B|}{|T|}
\]

\[
\text{Confidence}(A \Rightarrow B) = \frac{|A \cup B|}{|A|}
\]

Rules that pass both user-defined minimum support and minimum confidence threshold are called Strong Association Rules.

In this paper, we make use of the association rules technique in generating frequent item-sets in order to reveal the underlying patterns in wine profiles. Each wine review is considered as a transaction with the paired attributes acting as items of a transaction. As each review is processed, the attributes are recorded to build a collection of frequent wine descriptors. At this point in the algorithm, frequent item-sets would be used to find a correlation to a “label” in the transaction, thus finding when another item should be present based on the item-sets found. This so called “label” could be different extra information other than wine attributes, such as wine grade, region, grape type…etc. We choose wine grade as the label in this section. The goal is to accurately predict a wine is scored above 90 (classic and outstanding wines) or below 90(good and very good wines) based on only the sensory review.

We also provide the following example in Table 7 to show how we apply association rules to predict the range of the wine score. In the example, we have five wines in total, four of them with
known scores and we try to predict the score range (>90 or <=89) of the wine 5. Assume we define minimum support=50% and minimum confidence=80%. Based on wines 1~4, we can find one strong association rule (“CHERRY” and “BEAUTY” => “>90”) with support=2/4 and confidence=2/2. The rule indicates “if a wine has cherry and beauty in their review, it is a 90+ points wine”. Therefore, since this rule is applicable to the wine5, we then can predict it as a 90+ points wine.

Table 7. Example of wine score range prediction via association rules algorithm

<table>
<thead>
<tr>
<th>BLUEBERRY</th>
<th>CHERRY</th>
<th>CHEWY</th>
<th>TANNINS</th>
<th>BEAUTY</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wine 1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>95</td>
</tr>
<tr>
<td>Wine 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>85</td>
</tr>
<tr>
<td>Wine 3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>88</td>
</tr>
<tr>
<td>Wine 4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>91</td>
</tr>
<tr>
<td>Wine 5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>?</td>
</tr>
</tbody>
</table>

4.2 Dataset for Classification

Our first dataset is composed of the Top 100 wines in 2011, which means none of them is below 90 points. Therefore, the first dataset is not suitable for our classification experiment. We subscribed to an online Wine Spectator’s account and retrieved one thousand wine reviews with 250 in the [85-89] scores category, 250 in the [90-94] scores category and 250 in the [95-100] scores category.

Attributes were identified by scanning each wine review for terms found in the wine wheel dictionary, which is introduced in section 2. Each word in the review was compared against the “original terms” found in the wine wheel to verify the key descriptors used to denote wine flavors. Each recorded entry in the dataset was assigned the normalized name of the attribute. This normalized name ensured that small variations in how specific flavors were described were able to be combined as a single attribute of the wine. In this way, each wine in the dataset was represented by a collection of attributes that accurately describe the important flavors. These final attributes are then used as the basis in association rule mining for classifying the grade range of an input wine. Our second dataset with 1000 wines is also available under: http://www.cs.gsu.edu/~escbexc/Wine%20Informatics.htm

4.3 Results

We apply the association rules algorithm mentioned in section 4.1 on the dataset described in section 4.2. In order to obtain the authentic prediction accuracy, we use the 5-fold cross validation approach: The whole dataset is divided into five smaller datasets, each smaller dataset, which contains 100 wines with 90~100 scores and 100 wines with 80~89 scores, is treated as the testing dataset which the other 800 wines are treated as the training dataset. The training dataset will be responsible for generating association rules for classification. These rules will be tested on each wine. If no rule can be applied to the testing wine, we say this wine is unpredictable. If more than one rule can be applied to the testing wine, we use the rule with higher confidence value. The five prediction accuracy results generated from each smaller dataset are averaged to produce a single estimation.

In the association rules algorithm, users need to define the minimum support and confidence. Different user defined values will produce different results. In this experiment, the higher minimum support and confidence value, the more rules will be generated; thus, the more wines can possibly be predicted; however, the prediction accuracy may drop. Table 8 demonstrates an experimental 5-fold cross validation results based on minimum support=1% and minimum confidence=60%~90% with 10% interval.

In table 8, it is clear to see that with the same minimum support value, the best prediction accuracy (82.27%) is generated by minimum confidence=90%; however, the coverage is the lowest (61.90%); among 200 wines in the testing dataset, only 123.8 wines can be predicted. On the other side, the lowest prediction accuracy (72.86%) in the table is generated by minimum confidence=60% with 97.3% coverage.

We also provide our 5- fold cross validation results with different minimum support and confidence combination in table 9. The results also prove similar trend. Reading the results from top to bottom, each row represents the confidence parameter, marking the rate at which an item-set of descriptors shows up in association with one grade range of a wine. The more restrictive requirement of associations results in a higher accuracy and lower coverage due to the decreased amount of rules being generated.

Table 8. Experimental Results of prediction accuracy and coverage for 5-fold cross validation based on 1% minimum support and 60%, 70%, 80% and 90% minimum confidence

<table>
<thead>
<tr>
<th>Conf.</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>60%</td>
<td>70.98%</td>
<td>70.92%</td>
<td>68.56%</td>
<td>76.77%</td>
<td>77.08%</td>
<td>72.86%</td>
</tr>
<tr>
<td>70%</td>
<td>73.51%</td>
<td>71.20%</td>
<td>70.65%</td>
<td>77.84%</td>
<td>79.31%</td>
<td>74.50%</td>
</tr>
<tr>
<td>80%</td>
<td>73.72%</td>
<td>70.63%</td>
<td>74.85%</td>
<td>82.10%</td>
<td>79.22%</td>
<td>76.10%</td>
</tr>
<tr>
<td>90%</td>
<td>82.75%</td>
<td>79.39%</td>
<td>77.60%</td>
<td>87.20%</td>
<td>84.43%</td>
<td>82.27%</td>
</tr>
</tbody>
</table>

Table 9. 1% Support 5-Fold Accuracy

<table>
<thead>
<tr>
<th>Conf.</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>60%</td>
<td>96.50%</td>
<td>98.00%</td>
<td>97.00%</td>
<td>99.00%</td>
<td>96.00%</td>
<td>97.30%</td>
</tr>
<tr>
<td>70%</td>
<td>92.50%</td>
<td>95.50%</td>
<td>92.00%</td>
<td>92.50%</td>
<td>87.00%</td>
<td>91.90%</td>
</tr>
<tr>
<td>80%</td>
<td>78.00%</td>
<td>80.00%</td>
<td>85.50%</td>
<td>81.00%</td>
<td>77.00%</td>
<td>80.30%</td>
</tr>
<tr>
<td>90%</td>
<td>58.00%</td>
<td>65.50%</td>
<td>62.50%</td>
<td>62.50%</td>
<td>61.00%</td>
<td>61.90%</td>
</tr>
</tbody>
</table>
5. Conclusion
In this paper, we have accomplished the following tasks. (1) We propose a new data science area named Wine Informatics, which uses the wine savory reviews as the domain knowledge. (2) We proposed the computational wine wheel to retrieve attributes from wine reviews. (3) We show how these attributes can play an integral role in grouping different wines together through the clustering algorithm. (4) We develop a weighted similarity measure that can perform binary comparisons on our wine dataset. The clustering results generated by the hierarchical clustering algorithm suggest that when using only the attributes of a wine review, we can aggregate wines together that have similar world region, monetary value, vintage, type, and varietal. (5) We use the computational wine wheel to extract 1000 wine’s reviews and perform association rules algorithm on the generated dataset to predict whether or not a wine’s score is higher than 90. Considering the difficulty of processing the human savory reviews, we can aggregate wines together that have similar world region, monetary value, vintage, type, and varietal. (5) We use the computational wine wheel to extract 1000 wine’s reviews and perform association rules algorithm on the generated dataset to predict whether or not a wine’s score is higher than 90. Considering the difficulty of processing the human savory reviews and predicting the wine grade range, we are able to produce the accuracy results as high as 85.25%. We believe our prediction accuracy results are an achievement. It also proves that our prediction accuracy results are as high as 85.25%. We believe our prediction accuracy results are an achievement. It also proves that our prediction accuracy results are an achievement. It also proves that our prediction accuracy results are an achievement. It also proves that our prediction accuracy results are an achievement.

6. REFERENCES