Application of Data Mining Techniques to *Olea europaea var. media oblonga* production from Thassos Island

Theodosios Theodosiou[1], Stavros Valsamidis[1],Georgios Hatziliadis[1], Michael Nikolaidis[1]

[1]Department of Accountancy, Technology Educational Institute of Thessaloniki, Greece
e-mail: theodosios.theodosiou@gmail.com, {svalsam,mnikol}@teikav.edu.gr, egs-kav@otenet.gr

Abstract. A huge amount of data is produced in our days in the agriculture sector. Due to the huge amount of this datasets it is necessary to use data mining techniques in order to comprehend the data and extract useful information. In our work we apply three different data mining techniques to data about *Olea europaea var. media oblonga* from the island of Thassos, at the northern part of Greece. The data were from 1063 farmers from three different municipalities of Thassos, namely Kallirachi, Limenaria and Prinos and concerned 2010. They were analysed using the classification algorithm OneR, the clustering algorithm k-means and the association rule mining algorithm, Apriori from the WEKA data mining package. The results indicate that organic cultivation could improve the production of olives and olive oil.

Keywords: *Olea europaea*, Classification, Clustering, Association Rule Mining.

1 Introduction

Nowadays huge volumes of agricultural data are accumulated from a variety of sources on almost daily bases. The size of the datasets makes obsolete manual analysis and extraction of useful information. Consequently data mining techniques are becoming a necessity in order to process and analyse these datasets. Data Mining is an iterative process of creating predictive and descriptive models, by uncovering previously unknown trends and patterns in vast amounts of data, in order to extract useful information and support decision making (Mucherino et al., 2009).

The application of data mining techniques into research areas such as the agricultural field is quite an emerging area of research. The techniques applied to agriculture are not specifically designed for this field. On the contrary they are quite general and could be easily applied to any type of data. (Mucherino et al., 2009). Abdullah et al. (2004) applied data mining in order to discover rules for the use of

In: M. Salampasis, A. Matopoulos (eds.): Proceedings of the International Conference on Information and Communication Technologies
for Sustainable Agri-production and Environment (HAICTA 2011), Skiathos, 8-11 September, 2011.

pesticides in agriculture. Data mining methods are divided into three major categories. The first category involves the classification methods, whereas the second the clustering ones and the third the association rule mining methods.

Classification methods use a training dataset in order to estimate some parameters of a mathematical model that could in theory optimally assign each case from a new dataset into a specific class. In other words, the training set is used to train the classification technique how to perform its classification.

Clustering refers to methods where a training set is not available. Thus, there is no previous knowledge about the data to assign them to specific groups. In this case, clustering techniques can be used to split a set of unknown cases into clusters.

Association rule mining discovers relationships, sometimes hidden, among attributes (variables) in a dataset.

In our research we apply methods from these three data mining major categories to cultivation and production data of *Olea europea var. media oblonga* from the island of Thassos. Olive oil is a natural fruit juice with excellent nutritional characteristics. It is a typical lipid source food of the Mediterranean diet and its consumption has been associated with a low incidence of cardiovascular diseases, neurological disorders, breast and colon cancers, as well as with hipolipidemic and antioxidant properties. An increase of interest in olive oil as a healthy food has been observed lately in areas other than the Mediterranean countries mainly because of its fatty acid composition and content of other functional food components, such as polyphenols (Vekiari et al., 2010).

Olive cultivation is exceptionally spread in the island of Thassos, in Greece. The variety of olive cultivated all over Thassos is Throumbolia (Olea europaea var. media oblonga). The Throumbolia variety grows at altitudes of up to 700m and its fruits are medium-size. Its main characteristic is that under special conditions of temperature and moisture, the bitter taste, which is evident in the olives while they are still on the tree, disappears due to the hydrolysis of oleuropein by the action of the fungus *Phoma oleae*. It is important that olives are harvested as soon as they fall from the tree and pressed immediately to produce sweet oil, rich in aromatic substances, because of the olive trees height. Otherwise, a deterioration of olive quality usually occurs resulting in unpleasant tasting oil (Vekiari et al., 2010).

It is well known that the differences in olive oil quantity from various regions are attributed to olive variety, environmental factors, harvesting methods, time of harvest, and extraction techniques. Furthermore, there is little information on the specific factors that influence the variety of Throumbolia on the island of Thassos. Since, olive variety, harvesting methods, time of harvest and extraction techniques for public extraction factories are almost identical for the farmers in our study, we would like to investigate the difference in the environmental factors if and how they affect olive oil quantity. The trees are cultivated in three different areas of Thassos, namely Kallirachi, Limenaria and Prinos. Limenaria is considered to have a drier climate compared to the other two.

Furthermore, we also investigate the effects of organic cultivation or biological cultivation, as it is usually called in Greece, on the quantity of olive oil. Organic cultivation is the form of agriculture that relies on techniques such as crop rotation, green manure, compost and biological pest control to maintain soil productivity and control pests on a farm. Organic cultivation excludes or strictly limits the use of artificial fertilizers, pesticides (which include herbicides, insecticides and fungicides), plant growth regulators such as hormones, livestock antibiotics, food additives, and genetically modified organisms[1]. The latest years there are an increasing number of farmers that are changing their type of cultivation in the island of Thassos from conventional ones to organic.

Section 2 describes the dataset and the methodology used in our research. Sections 3 presents the results from the data mining methods and Section 4 discusses the results and refers to the main conclusions of our research.

2. Data and Methodology

In this section the dataset we used in our methodology is described in detail. Also, the data mining methods applied to the olive oil data are explained and analysed.

2.1 The Dataset

The dataset was collected from the Enosi Agrotikon Sinetairismon (EAS) of Kavala[2]. The data were collected during 2010 and involve 1063 farmers and 3 different municipalities from the island of Thassos, namely Kallirachi, Limenaria and Prinos. Prinos and Kallirachi are at the northern part of Thassos whereas Limenaria at the southern part. The data are originally in ASCII form and are obtained from the EAS data repository. Each farmer is described by 8 variables in the repository of EAS. The first two variables are the AFM (tax id) and the last name of the farmer and they are omitted from our analysis, since they do not influence the production of olive oil and are private data. The remaining six variables are used in the analysis described in the methodology section and are called Area, Trees, Bio, Mun, Olives and Oil. Table 1 describes each variable in detail.

Table 1. The variables used in our analysis

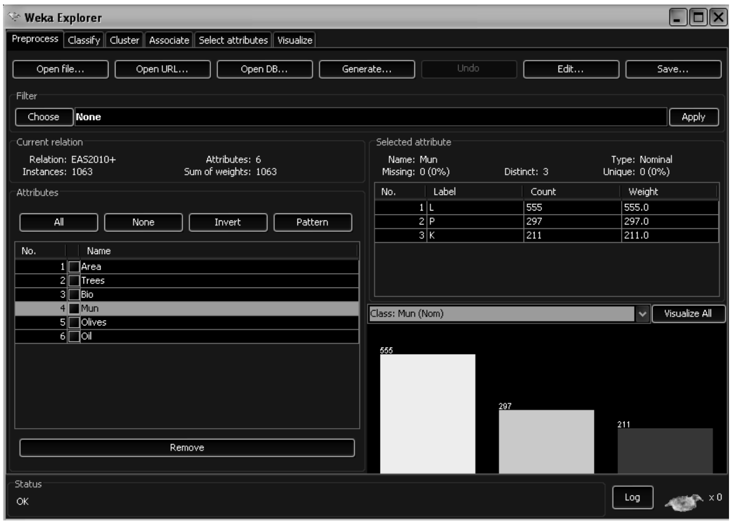
Variable Name	Description	Type
Area	The total area of land owned by the farmer in m^2	Numeric
Trees	The total number of trees owned by a farmer	Numeric
Bio	The type of cultivation (0: Typical, 1: Biological/Organic)	Numeric
Mun	The municipality in which the farmer has his trees	Nominal
Olives	The number of olives each farmer has	Numeric
Oil	The total liters of olive oil produced per farmer	Numeric

[1] Directorate General for Agriculture and Rural Development of the European Commission, http://ec.europa.eu/agriculture/organic/organic-farming/what-organic_en

[2] EAS Kavalas - http://www.easkavalas.gr/

2.2 Data mining methods

The WEKA (*Waikato Environment for Knowledge Analysis*) (Witten & Frank, 2005) computer package was used in order to apply classification, clustering and association rule mining methods to the dataset. WEKA is open source software that provides a collection of machine learning and data mining algorithms. Fig. 1 shows the basic Graphical User Interface (GUI) of WEKA. One of the main objectives of WEKA is to mine information from existing agricultural datasets (Cunningham and Holmes, 1999) and the main reason for choosing it for analyzing our data.



Filter

None

Current relation

Relation: EAS2010+ Attributes: 6
Instances: 1063 Sum of weights: 1063

Attributes

No.		Name
1	□	Area
2	□	Trees
3	□	Bio
4	■	Mun
5	□	Olives
6	□	Oil

Selected attribute

Name: Mun Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	L	555	555.0
2	P	297	297.0
3	K	211	211.0

Class: Mun (Nom)

Status

OK

Fig. 1. WEKA environment

There are various classification methods implemented in WEKA, like Naïve Bayes, ZeroR, OneR, etc. In the *classification* step, the algorithm *OneR* (Witten & Frank, 2000) was applied to our data. It uses the minimum-error attribute for prediction, discretizing numeric attributes (Holte, 1993). The main advantage is that it produces very simple rules for classification and can be considered the baseline for classification performance. It was found to perform as well as more sophisticated algorithms when applied to many of the standard machine learning test datasets (Holte, 1993). OneR can parsimoniously discover and represent simple relationships between the real data (Cunningham and Holmes, 1999). In our case the variable “mun” which shows the municipality the trees are cultivated in is used as a class.

The *clustering* step uses the k-means algorithm (MacQueen, 1967; Kaufmann & Rousseeuw, 1990), called *SimpleKMeans* in WEKA. K-means is an efficient partitioning algorithm that decomposes the data set into a set of k disjoint clusters. It is a repetitive algorithm in which the items are moved among the various clusters until they reach the desired set of clusters. With this algorithm a great degree of similarity for the items of the same cluster and a large difference of items, which belong to different clusters, are achieved. The Euclidean distance is used to compute the differences between the olive trees cultivations. The variable “bio” is used in order to assess the accuracy of the clustering and investigate its impact on olive tree cultivation.

Association rule mining is one of the most well studied data mining tasks. It discovers relationships among attributes (variables) in datasets, producing if-then statements concerning attribute-values (Agarwal, Imielinski, & Swami, 1993). An association rule $X \Rightarrow Y$ expresses a close correlation among items in a dataset, in which transactions in the dataset where X occurs, there is a high probability of having Y as well. In an association rule X and Y are called respectively the antecedent and consequent of the rule. The strength of such a rule is measured by values of its support and confidence. The *confidence* of the rule is the percentage of transactions with antecedent X in the dataset that also contain the consequent Y. The *support* of the rule is the percentage of transactions in the dataset that contain both the antecedent and the consequent Y in all transactions in the dataset.

The WEKA system has several association rule-discovering algorithms available. The Apriori algorithm (Agarwal et al., 1996) is used for finding association rules over the discretized LMS data table in Appendix 1. Apriori (Agrawal, 1994) is the best-known algorithm to mine association rules. It uses a breadth-first search strategy to counting the support of item sets and uses a candidate generation function, which exploits the downward closure property of support. Iteratively reduces the minimum support until it finds the required number of rules with the given minimum confidence.

There are different techniques of categorization for association rule mining. Most of the subjective approaches involve user participation in order to express, in accordance with his/her previous knowledge, which rules are of interest. One technique is based on unexpectedness and actionability (Liu et al, 1996; Liu et al, 2000). *Unexpectedness* expresses which rules are interesting if they are unknown to the user or contradict the user’s knowledge. *Actionability* expresses that rules are interesting if users can do something with them to their advantage. The number of rules can be decreased to unexpected and actionable rules only (García et al., 2008). Another technique proposes the division of the discovered rules into three categories (Minaei-Bidgoli et al., 2004). (1) *Expected and previously known*: This type of rule confirms user beliefs, and can be used to validate our approach. Though perhaps already known, many of these rules are still useful for the user as a form of empirical verification of expectations. For agriculture, this approach provides opportunity for rigorous justification of many long held beliefs. (2) *Unexpected*: This type of rule contradicts user beliefs. This group of unanticipated correlations can supply interesting rules, yet their interestingness and possible actionability still requires further investigation. (3) *Unknown*: This type of rule does not clearly belong to any category, and should be categorized by domain specific experts.

3. Results

The first step before applying the data mining methods described in the previous section is the pre-processing of the data in order to prepare them for data analysis.

3.1 Pre-processing

The three municipalities in the variable "mun", namely Kallirachi, Limenaria and Prinos were replaced with the values K, L and P respectively. Furthermore, certain filters were applied to the data, such as the filter *NumericalToNominal* in order to convert numeric variables and their values to nominal. For example, number 0 and 1 in variable Bio are converted to nominal, where 0 signifies conventional cultivation and 1 organic.

Fig. 2 depicts all the variables used in our analysis. The different scales of gray correspond to the three different municipalities. Light gray corresponds to Kallirachi, medium gray to Prinos and dark gray (black) to Limenaria. One can see that 555 farmers where from the area of Limenaria, 297 from Prinos and 211 from Kallirachi. Also it is evident that for variables area, trees, olives and oil the distribution of their values is right-skewed. It is noteworthy that only 197 farmers use organic cultivation and that most of them are in the area of Limenaria.

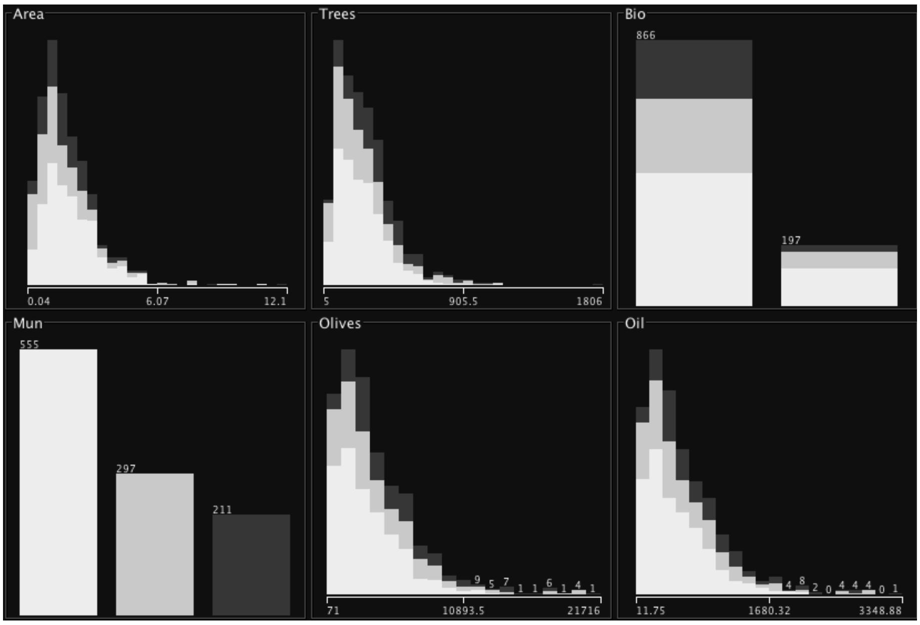


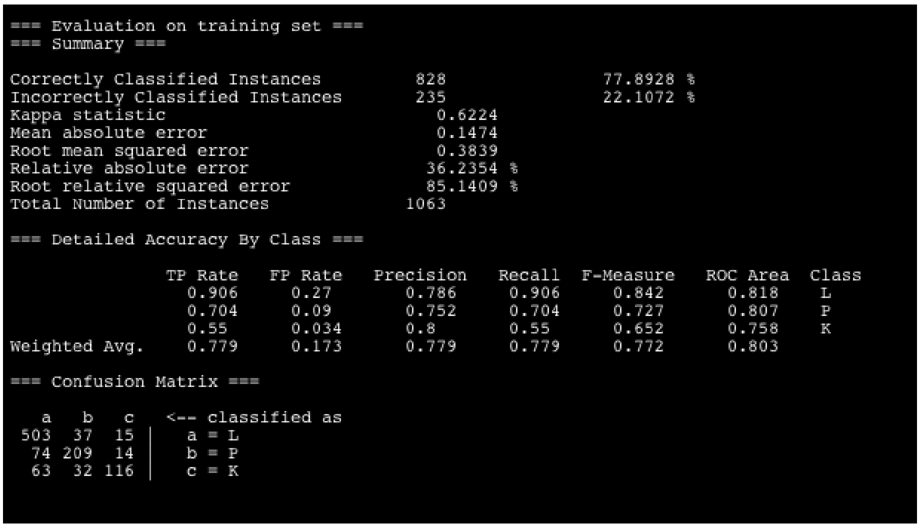
Fig. 2. Visualization of the attributes with class variable "Mun" (municipality)

3.2 Classification

In the classification step, the algorithm *OneR* is applied. The attribute "Mun" (municipality) is used as a class. Fig. 3 presents the overall accuracy of the model computed from the training dataset and is equal to 77.89%. The worst performance based on the F-measure that combines precision and recall is for the municipality of

Kallirachi and equals 65.2%, whereas the best performance is for the area of Limenaria and equals 84.2%.

The results indicate that the best attribute, which describes the classification, is variable Olives that holds the number of olives for each farmer. This means that variable Olives is more closely related to variable Mun than the other variables and therefore in some Mun (municipalities) there is higher olives production (Kallirachi, Prinos) than in other municipalities (Limenaria). A possible explanation for these results is that the area of Limenaria has a dryer climate compared to Prinos and Kallirachi and thus the olive trees have lower production.



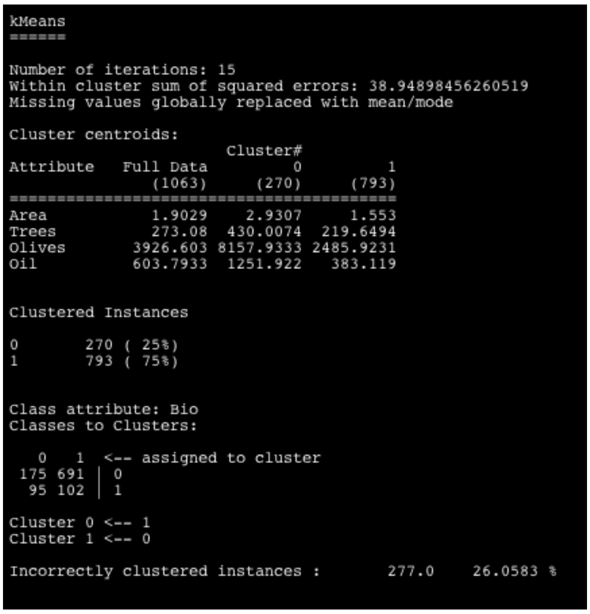
```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         828               77.8928 %
Incorrectly Classified Instances       235               22.1072 %
Kappa statistic                          0.6224
Mean absolute error                      0.1474
Root mean squared error                  0.3839
Relative absolute error                 36.2354 %
Root relative squared error             85.1409 %
Total Number of Instances             1063

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.906     0.27       0.786     0.906     0.842      0.818    L
                 0.704     0.09       0.752     0.704     0.727      0.807    P
                 0.55      0.034      0.8       0.55      0.652      0.758    K
Weighted Avg.    0.779     0.173      0.779     0.779     0.772      0.803

=== Confusion Matrix ===

   a   b   c   <-- classified as
 503  37  15 |   a = L
  74 209  14 |   b = P
  63  32 116 |   c = K
```

Fig. 3. Classification results using variable “Mun” (municipality) as class

3.3 Clustering

The clustering step was performed using the k-means algorithm (*SimpleKmeans* in the context of WEKA). The number of clusters is set to 2, since the variable “bio” was used to compute the accuracy of the clustering and inspect the impact of the type of the cultivation to olive oil production.

Fig. 4 shows the results of the clustering. The incorrectly clustered instances (farmers) are 26.06% based on variable “bio”. It is also evident from the cluster centroids that organic cultivation, represented as cluster 0 in the results, contains trees that produce higher quantities of olive oil compared to cluster 1 (conventional cultivation). This preliminary result indicates the potential of organic cultivation for improving olive oil production.



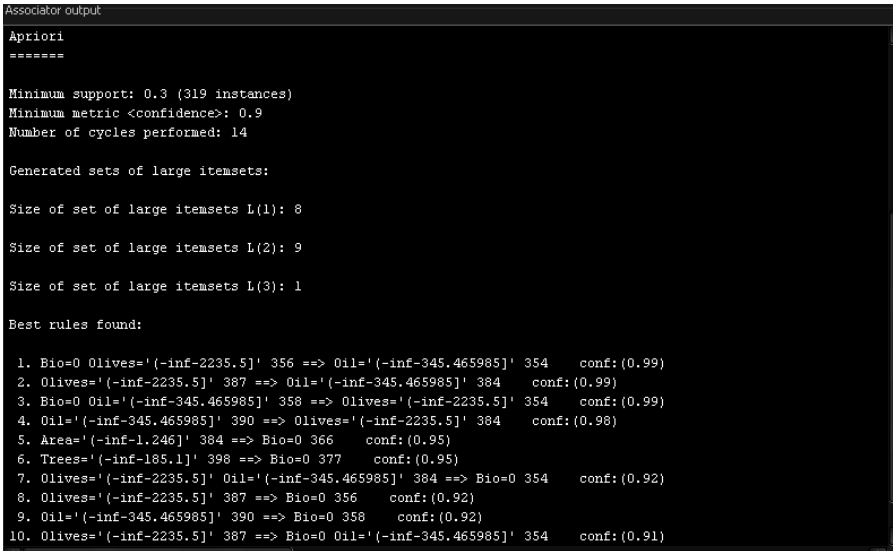
```
kMeans
======

Number of iterations: 15
Within cluster sum of squared errors: 38.94898456260519
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
                (1063)      (270)      (793)
============================================
Area            1.9029     2.9307      1.553
Trees           273.08   430.0074   219.6494
Olives        3926.603  8157.9333  2485.9231
Oil           603.7933   1251.922    383.119


Clustered Instances

0        270 ( 25%)
1        793 ( 75%)


Class attribute: Bio
Classes to Clusters:

   0   1  <-- assigned to cluster
 175 691 | 0
  95 102 | 1

Cluster 0 <-- 1
Cluster 1 <-- 0

Incorrectly clustered instances :       277.0    26.0583 %
```

Fig. 4. Clustering results. Variable “bio” is used for assessing the clustering.

3.4 Association rule mining

The Apriori algorithm (Agarwal et al., 1996) was used for finding association rules for our dataset. The algorithm was executed using a minimum support of 0.1 and a minimum confidence of 0.9, as parameters. WEKA produced a list of 10 rules (Fig. 5) with the support of the antecedent and the consequent (total number of items) at 0.1 minimum, and the confidence of the rule at 0.9 minimum (percentage of items in a 0 to 1 scale).

The application of the Apriori algorithm for association provided some interesting outcomes for the production of olive trees. Figure 5 shows the association rules that can be discovered. There are of course some uninteresting rules, like rules 2 and 4. They present relatively known information since it is an expected or conforming relationship between variables Olives and Oil. There are also a couple of symmetrical rules, since the antecedent element and the consequent element are interchanged. There is also a similar triad of rules, rules with the same element in antecedent and consequent but interchanged, such as rules 1, 3 and 7. It is rather surprising the outcome of rules 5, 6, 7, 8 and 9. Conventional (non-organic) is the preferable cultivation for farmers with small area (and trees), and consequently small production of olives and olive oil. The last remark might be useful for the motivation that might be given to farmers with limited number of trees (or area) to follow the organic cultivation.



```
Apriori
=======

Minimum support: 0.3 (319 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 14

Generated sets of large itemsets:

Size of set of large itemsets L(1): 8

Size of set of large itemsets L(2): 9

Size of set of large itemsets L(3): 1

Best rules found:

 1. Bio=0 Olives='(-inf-2235.5]' 356 ==> Oil='(-inf-345.465985]' 354    conf:(0.99)
 2. Olives='(-inf-2235.5]' 387 ==> Oil='(-inf-345.465985]' 384    conf:(0.99)
 3. Bio=0 Oil='(-inf-345.465985]' 358 ==> Olives='(-inf-2235.5]' 354    conf:(0.99)
 4. Oil='(-inf-345.465985]' 390 ==> Olives='(-inf-2235.5]' 384    conf:(0.98)
 5. Area='(-inf-1.246]' 384 ==> Bio=0 366    conf:(0.95)
 6. Trees='(-inf-185.1]' 398 ==> Bio=0 377    conf:(0.95)
 7. Olives='(-inf-2235.5]' Oil='(-inf-345.465985]' 384 ==> Bio=0 354    conf:(0.92)
 8. Olives='(-inf-2235.5]' 387 ==> Bio=0 356    conf:(0.92)
 9. Oil='(-inf-345.465985]' 390 ==> Bio=0 358    conf:(0.92)
10. Olives='(-inf-2235.5]' 387 ==> Bio=0 Oil='(-inf-345.465985]' 354    conf:(0.91)
```

Fig. 5. The Apriori algorithm results based on the confidence metric

Summarizing the results from the classification, the clustering and the association rule mining methods we can conclude that:

(i) The attribute which best describes the classification is the variable Olives that holds the number of olives for each farmer. The attribute "Mun" (municipality) is used as a class.

(ii) Using "bio" as class attribute in clustering, namely if cultivation is conventional or organic, the results show that trees which belong to the second cluster produce higher quantities of olive and oil compared to the ones with conventional cultivation.

(iii) In association rule mining, although there are some trivial rules, namely expected and previously known, like rules 2 and 4 that show that the productions of oil and olives are mutually dependent, there are also rules like 1, 3, 8, 9 and 10 which offer a lot of actionability. Organic cultivation has great impact to the olives and oil production. Rules 5 and 6 show that farmers with limited number of trees and/or area prefer the organic cultivation.

Overall, the production of olives and oil are based on two different factors. The first one is environmental and depends on the part of the island that the olives trees are cultivated. The second factor is the type of cultivation, traditional or organic.

4. Conclusions

The application of three different data mining techniques in agricultural data is presented in this paper. The dataset involves *Olea europea var. media oblonga* that is cultivated in the island of Thassos. It is quite a commercially successful *Olea* species that the latest years has been extensively monitored by the EAS of Kavala. The parameters which were investigated were the total area of land owned by the farmer in m^2, the total number of trees owned by a farmer, the type of cultivation (0: Typical, 1: Biological/Organic), the municipality in which the farmer has his/her trees, the number of olives each farmer has and the total liters of olive oil produced per farmer.

The results show some interesting outcomes. First of all, the classification method indicate that the environmental differences between the northern and southern part of the island can influence the production of olives and oil, concerning the number of olives and the quantities of olive oil they can produce. This is mainly due to the fact that the southern part of the island is considered to be dryer than the northern part and thus the amount of waterfall influences olive trees. Furthermore, the clustering results show that organic cultivation could improve olive and oil production. Although, organic cultivation is usually more cumbersome than conventional one, it could worth the extra effort if olive oil production is increased. The last data mining method, the association rule mining, suggests that farmers with limited number of trees (or area) could benefit more by following the organic cultivation. Organic cultivation could help them produce more olives and olive oil. In other words organic cultivation seems to help improve production in the regions presented in the study at the island of Thassos.

The study has an inherent advantage as far as the data analysis is concerned. The tree *Olea* offers two variables as input, the total number of trees and the total area of land and two variables as output, the number of olives and the total volume of olive oil.

Further investigation is still required since the results are based on only three municipalities. In the future we plan to extend the study to other municipalities and areas of the island Thassos. Furthermore, we plan to overcome the limitation of manually mining such datasets, by developing a plug-in tool for WEKA to automate the whole procedure. It should be mentioned that even if the scope of the method is on olive trees production, it could be easily adopted to the production of other products of agriculture. A comparative analysis for the same products among different parts of Greece could also be useful, i.e. for *Olea* among Thassos and Lesvos islands and Chalkidiki peninsula.

References

1. Abdullah, A., Brobst, S., Pervaiz, I., Umer, M., Nisar, A. (2004). Learning dynamics of pesticide abuse through data mining, Proceeding ACSW Frontiers '04 Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation - Volume 32.
2. Darcy Miller, Jaki McCarthy, Audra Zakzeski, (2009) A Fresh Approach to Agricultural Statistics: Data Mining and Remote Sensing, Section on Government Statistics – JSM 2009

3. Agrawal, R. & Srikant, R. (1994). Fast algorithms for mining association rules. Proceedings of 20th International Conference on Very Large Data Bases (pp. 487-499).
4. Holte, R.C. (1993). Very simple classification rules perform well on most commonly used datasets. Machine Learning. 11,63-91.
5. Kaufmann, L., Rousseeuw, P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis, New York, John Wiley & Sons.
6. Liu, B. & Hsu, W. (1996). Post-Analysis of Learned Rules. Proceedings of National Conference on Artificial Intelligence. Portland, Oregon, USA, (pp. 828–834).
7. Liu, B., Hsu, W., Chen, S. & Ma, Y. (2000). Analyzing the Subjective Interestingness of Association Rules. IEEE Intelligent Systems, 15(5), 47–55.
8. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth berkeley symposium on mathematical statistics and probability, (pp. 281–297). California, USA.
9. Minaei-Bidgoli, B., Tan, P-N. & Punch, W.F. (2004). Mining Interesting Contrast Rules for a Web-based Educational System. Proceedings of Int. Conf. on Machine Learning Applications, Louisville, USA 2004 (pp. 320- 327).
10. Mucherino, A., Papajorgji, P., Pardalos, P.M. (2009). A survey of data mining techniques applied to agriculture. Oper Res Int J, DOI 10.1007/s12351-009-0054-6.
11. Cunningham, S. J., and Holmes, G. (1999). Developing innovative applications in agriculture using data mining. In the Proceedings of the Southeast Asia Regional Computer Confederation Conference.
12. Vekiari, S.A., Oreopoulou, V., Kourkoutas, Y., Kamoun, N., Sallem, M., Psimouli, V., Arapoglou, D. (2010). Characterization and seasonal variation of the quality of virgin olive oil of the Throumbolia and Koroneiki varieties from Southern Greece. Grasas Y Aceites, 61 (3), pp. 221-231.
13. Witten, I. & Frank, E., (2005). Data Mining Practical Machine Learning Tools and Techniques, San Francisco: Morgan Kaufmann.