SOFTWARE SOLUTIONS IN MARKETING RESEARCH FOR KNOWLEDGE DISCOVERY IN DATABASES BY FUZZY CLUSTERING

SOFTVERSKA RJEŠENJA U MARKETING ISTRAŽIVANJU ZA OTKRIVANJE ZNANJA U BAZAMA PODATAKA POMOĆU FUZZY KLASTERIRANJA

Brano Markić, Dražena Tomić
Faculty of Economics, University of Mostar, Mostar, Bosnia and Herzegovina
Ekonomski fakultet, Sveučilište u Mostaru, Mostar, Bosna i Hercegovina

Abstract

Knowledge discovery in databases is the process of identifying novel, valid, useful and ultimately understandable patterns in data stored in databases. Data mining is only a step in this process in charge to find patterns or models in data. There are many data mining algorithms for clustering. Clustering is unsupervised classification, the process of grouping the data into classes so that the data objects (examples) are similar to one another within the same cluster and dissimilar to the objects in other clusters. In the paper is developed a conceptual model and program solution for clustering data stored in subject-oriented data warehouse. Data warehouse and mining algorithms are integrated and this integration has shown satisfactory implementation power.

1. INTRODUCTION

Knowledge discovery in databases is an interdisciplinary field that relates to databases, data warehouse, machine learning, expert systems (formalisms of knowledge representation), statistics and operational research and data visualization. The common goal of integrating these different fields is extracting knowledge from data stored in large databases and data warehouse. For us is especially interesting the overlap knowledge data discovery and the area of data warehouse. The most popular approach for analysis of data warehouses is OLAP (on-line analytical processing) and is focusing on providing multi-dimensional data analysis. Knowledge discovery and OLAP are complementary and may be integrated with other technologies such as expert systems and provide a new generation of intelligent information extraction and management tools.

2. KNOWLEDGE DISCOVERY IN DATABASES FOR MARKETING RESEARCH

The process of knowledge discovery inherently consists of several steps [1]. The first step is to understand the application domain and to formulate the problem. Our application domain is clustering customers on business markets. This step determines the appropriate data mining algorithms. The second step is to collect and preprocess the data. Corresponding data model is physically implemented by relational database. This database is known as operational. From operational database in a process of extracting, transforming and loading data is populated the data warehouse. This step is one of the most tedious and usually takes the most time needed for the whole knowledge data discovery process.
Data for customers clustering are stored in relational database that is temporarily loading from transactional data bases. The fact table in data warehouse named as Customers_Cluster includes three measures: sales, profit and days of payments.

![Data warehouse for customers clustering](image)

The third step is data mining that extracts patterns or models hidden in data. A model can be viewed “a global representation of a structure” and “a pattern is a local structure, perhaps relating to just a handful of variables and a few cases”. In our example the third step is implementing fuzzy c-means clustering in marketing that extracts patterns hidden in data (adequate centroids of clusters).

The last step is to interpret discovered knowledge and put knowledge in practical use. All these steps are “covered” with adequate software technologies and programs. In the paper are shown the most important and critical steps in building the software solution of knowledge discovery.

### 3. SOFTWARE SOLUTION OF CLUSTERING IN MARKETING FUNCTIONAL FIELD

The software solution for market segmentation uses the fuzzy c-means algorithm. Clustering is the process of grouping the data into classes (clusters) so that the data objects (examples) are similar to one another within the same cluster and dissimilar to the objects in other clusters. A good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity.

Hard k-means algorithm executes a sharp clustering, in which each object is either assigned to a cluster or not.

The k-means algorithm partitions a set of N vector into c clusters (clusters $G_j, j=1,..,c$). The goal is finding cluster centers (centroids) for each cluster. The algorithm minimizes a dissimilarity (or distance) function which is given in Equation 1.

$$ J = \sum_{i=1}^{c} \sum_{x \in G_i} d(x_i - c_j) $$

$c_j$ is the centroid of cluster $i$; $d(x_i - c_j)$ is the distance between $x_i$ and $c_j$.

Overall dissimilarity function is expressed as in Equation 2.

$$ J = \sum_{i=1}^{c} \sum_{x \in G_i} \alpha_i (x - c_j)^2 $$

Partitioned groups can be defined by a binary membership matrix $(m)$, where the element $m_{ij}$ is 1 if the $x_i$ data point belongs to cluster $j$, and 0 otherwise (Equation 3)

$$ m_{ij} = \begin{cases} 1 & \text{if } \|x_i - c_j\|^2 < \|x_i - c_k\|^2, \text{for each } k \neq j, \\ 0 & \text{otherwise} \end{cases} $$

Since a record can only be in a cluster, the membership matrix $(m)$ has two conditions which are given equation 4 and equation 5.

$$ m_{ij} = \begin{cases} 1 & \text{if } \|x_i - c_j\|^2 < \|x_i - c_k\|^2, \text{for each } k \neq j, \\ 0 & \text{otherwise} \end{cases} $$

Centroids are computed as the mean of all vectors in group $i$:

$$ c_j = \frac{1}{|G_i|} \sum_{x_i \in G_i} x_i $$

The software solution for k-means algorithm and all necessary steps [2] no guarantee for an optimum solution steps with a recordset $x_i, j=1,..,n$. The performance of the algorithm depends on the initial positions of centroids.

Fuzzy clustering allows that one tuple belongs at the same time to several clusters but with different degrees. This is an important feature for segmentation business markets to increase the sensitivity.

Fuzzy c-means clustering is a clustering technique which is separated from hard k-means that employs hard partitioning. The employs fuzzy partitioning such that a tuple (fact table record in data warehouse) can belong to all groups with different membership grades between 0 and 1. Fuzzy c-means is an iterative algorithm. The aim of fuzzy c-means is to find cluster centers (centroids) that minimize a dissimilarity function.

1. The first step is randomly generating the clusters. The clusters are chosen from fact table DataKlasterF and the next source code written in Visual Basic shows how is randomly selecting these seed clusters.
2. The second step is calculating the membership matrix \( U \) according to Equation 7

\[
\hat{\alpha}_i u^i_j X_i
\]

\[
\hat{\alpha}_i u^i_j = \frac{1}{\hat{\alpha}_i \frac{d_{ij}}{d_j}}^2
\]

The next code written in Visual Basic implements the equation 10 (membership matrix \( m(i,j) \)):

```vba
rsDataKlasterF.MoveFirst ' rsDataKlasterF is a recordset
ReDim Preserve m(rc, Val(Trim$(Text1.Text)))
ReDim Preserve a(Val(Trim$(Text1.Text)))
For i = 1 To rc
    X = rsDataKlasterF!promet, Y = rsDataKlasterF!ruc,
    Z = rsDataKlasterF!dop
    rsCentroidF.MoveFirst
    For j = 1 To Val(Trim$(Text1.Text))
        xc = rsCentroidF!promet
        yc = rsCentroidF!ruc
        zc = rsCentroidF!dop
        a(j) = Round(Round((X - xc) ^ 2, 0) +
            Round((Y - yc) ^ 2, 0) + (Z - zc) ^ 2, 0)
        rsCentroidF.MoveNext
    Next j
    Dim S As Single
    For k = 1 To Val(Trim$(Text1.Text))
        If a(k) = 0 Then
            m(i, k) = 1
            S = 0
            Exit For
        Else
            S = S + Round(a(j) / a(k), 5)
        End If
    Next k
    If S <> 0 Then
        m(i, j) = Round(1 / S, 5)
    Else
        m(i, j) = 0
    End If
End If
```

The dissimilarity function which is used in fuzzy c-means clustering is given Equation 8

\[
J(U, c_1, c_2, ..., c_n) = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{m_{ij}}{d_{ij}}
\]

\( m_{ij} \) is between 0 and 1; \( c_i \) is the centroid of cluster \( i \); \( d_{ij} \) is the Euclidian distance between \( i \)th centroid \( (c_i) \) and \( j \)th data point; \( t \in [1,\infty] \) is a weighting exponent.

To reach a minimum of dissimilarity function there are two conditions. These are given in Equation 9 and Equation 10.

\[
f_j = \frac{\sum_{i=1}^{n} m(i, j)^t}{\sum_{j=1}^{m} (m(i, j))^t}
\]

\[
m(i, j) = \frac{1}{\sum_{j=1}^{m} \frac{d_{ij}}{d_j}}
\]

The membership matrix \( m(i,j) \) for the first fourteen records is presented by Fig. 2:

3. The third step is calculating the center of centroids (new centroids of clusters) using membership matrix \( m(i,j) \) and values for profits, sales and days of payments:

\[
c_j = \frac{\sum_{i=1}^{n} m(i, j)^2 \cdot X_i}{\sum_{i=1}^{n} m(i, j)^2}
\]

\( m(i,j) \) is between 0 and 1; \( c_j \) is the centroid of cluster \( j \). After calculating the new centroid it is necessary to calculate the distance between the cluster center and each one records. Stop if its improvement over previous iteration is below a threshold. Stop condition in this example is defined by statement:

```vba
If Abs(pC(j, 1) - rsCentroidF!promet) < 60 And
Abs(rucC(j, 2) - rsCentroidF!ruc) < 6 And
Abs(dopC(j, 3) - rsCentroidF!dop) < 1 Then
    nastavi = False
Else
    nastavi = True
End If
```

The next code implements calculating new centroids, test the conditions and if the stop condition is not satisfied calculate the new membership matrix.
Dim Brojnik() As Single, Nazivnik()
Dim nastavi As Boolean
Dim Xr() As Single
ReDim Xr(Val(Trim(Text1.Text)), 3)
ReDim Brojnik(Val(Trim(Text1.Text)), 3)
ReDim Nazivnik(Val(Trim(Text1.Text)), 3)
Dim p As Single, ruc, dop
Dim pC() As Single
Dim rucC() As Single
Dim dopC() As Single
ReDim pC(Val(Trim(Text1.Text)), 3)
ReDim rucC(Val(Trim(Text1.Text)), 3)
ReDim dopC(Val(Trim(Text1.Text)), 3)

Dim pC() As Single
Dim rucC() As Single
Dim dopC() As Single

nastavi = True
Do While nastavi
    rsCentroidF.MoveFirst
    For j = 1 To Val(Trim(Text1.Text))
        Brojnik(j, 1) = 0, Nazivnik(j, 1) = 0, Brojnik(j, 2) = 0,
        Nazivnik(j, 2) = 0
        Brojnik(j, 3) = 0, Nazivnik(j, 3) = 0
    Next j
    rsDataKlasterF.MoveFirst
    For i = 1 To rc
        Xr(j, 1) = rsDataKlasterF!promet, Xr(j, 2) = rsDataKlasterF!ruc
        Xr(j, 3) = rsDataKlasterF!dop
        Brojnik(j, 1) = Brojnik(j, 1) + Round(m(i, j) ^ 2 * Xr(j, 1), 2)
        Nazivnik(j, 1) = Round(Nazivnik(j, 1) + m(i, j) ^ 2, 3)
        Brojnik(j, 2) = Round(Brojnik(j, 2) + m(i, j) ^ 2 * Xr(j, 2), 3)
        Nazivnik(j, 2) = Round(Nazivnik(j, 2) + m(i, j) ^ 2, 3)
        Brojnik(j, 3) = Round(Brojnik(j, 3) + m(i, j) ^ 2 * Xr(j, 3), 3)
        Nazivnik(j, 3) = Round(Nazivnik(j, 3) + m(i, j) ^ 2, 3)
    Next i
    rsDataKlasterF.MoveNext
    For i = 1 To rc
        Xr(j, 1) = rsDataKlasterF!promet, Xr(j, 2) = rsDataKlasterF!ruc
        Xr(j, 3) = rsDataKlasterF!dop
        Brojnik(j, 1) = Brojnik(j, 1) + Round(m(i, j) ^ 2 * Xr(j, 1), 2)
        Nazivnik(j, 1) = Round(Nazivnik(j, 1) + m(i, j) ^ 2, 3)
        Brojnik(j, 2) = Round(Brojnik(j, 2) + m(i, j) ^ 2 * Xr(j, 2), 3)
        Nazivnik(j, 2) = Round(Nazivnik(j, 2) + m(i, j) ^ 2, 3)
        Brojnik(j, 3) = Round(Brojnik(j, 3) + m(i, j) ^ 2 * Xr(j, 3), 3)
        Nazivnik(j, 3) = Round(Nazivnik(j, 3) + m(i, j) ^ 2, 3)
    Next i
    rsDataKlasterF.MoveNext
    For i = 1 To rc
        X = rsDataKlasterF!promet
        Y = rsDataKlasterF!ruc
        Z = rsDataKlasterF!dop
        xc = rsCentroidF!promet
        yc = rsCentroidF!ruc
        zc = rsCentroidF!dop
        a(j) = Round(Round((X - xc) ^ 2, 0) +
                     Round((Y - yc) ^ 2, 0) +
                     Round((Z - zc) ^ 2, 0))
        rsCentroidF.MoveNext
        If j = 1 To Val(Trim$(Text1.Text))
            S = 0
            For k = 1 To Val(Trim$(Text1.Text))
                If a(k) = 0 Then m(i, k) = 1
                Exit For
            Next k
            If S <> 0 Then m(i, j) = Round(1 / S, 5)
            S = S + Round(a(j) /
                S = S + Round(a(j) /
        Next j
    Next i
End Loop

By iteratively updating the cluster centers and the membership matrix [3] for each record, fuzzy c-means iteratively moves the cluster centers to the “right center” within data records.

Cluster centers (centroids) are initializing using randomly selecting records and fuzzy c-means does not ensure that it converges to an optimal solution. Performance depends on initial centroids and may be improved on two ways:

1) Using an algorithm to determine all of the centroids. (for example: arithmetic means of all records)
2) Run fuzzy c-means several times each starting with different initial centroids.

Software agents for market segmentation preferred the second approach.

4. EXPERIMENTAL RESULTS

Software agent for segmentation business markets assigns the customers to different clusters, different market segments. After final segmentation of business markets to six segments and assign each one customer two segments follows six clusters with centroids and corresponding membership grades:
Now is necessary the expert judgment to assign the customers 6, 17, 83, 18, 67, 86, 88, 89 and 91 to adequate market segment. Namely, the software solution extracts the customers for which the membership grade is between 0.4 and 0.6. This interval may be closer.

5. CONCLUSION

These paper theoretically and practically presented the integration knowledge discovery in databases, data mining and marketing research. The data mining component of the knowledge discovery process is mainly concerned with algorithms by which patterns are extracted from the data (fuzzy c-means clustering). Data for customers clustering are stored in relational data warehouse that is temporarily loading from transactional data bases. After running the program for fuzzy c-means clustering written in Visual Basic development environment follow the results that is easy understand and explain. Finally step is finding out suspicious records, customers for which the membership degree is close to 0.5. Expert judgment is necessary to analyze this tuples. On the business market the firm may very easy to define the required number of clusters (five, ten, twenty etc.) and the software will assign the customers to adequate cluster with membership degree. So has been assured the sensitivity and broad applicability the software and the concept of knowledge discovery.

Literature

2. G. Berks, D.G. Keyserlingk, J. Jantzen, M. Dotoli, H. Axer, Fuzzy Clustering- A Versatile Mean to Explore Medical Database, ESIT2000, Aachen, Germany

Fig. 3. Clusters centroids and corresponding membership grades

These software solution shows that the first customer (record 1) 93.913% belongs to market segment where the average amount of sale is 2446.555; average number of payments day is 50 and realized profit per customer 324.259.

One customer can belong to several market segments at the same time but with different degrees. This is an important feature for market segmentation to increase the sensitivity.

For the membership degrees close to 0.5 are the suspicious cases. Assigning the sample to one cluster could be wrong. Therefore software gives very reliable results.

Fig. 4.: Membership degrees of suspicious samples

Now is necessary the expert judgment to assign the customers 6, 17, 83, 18, 67, 86, 88, 89 and 91 to adequate market segment. Namely, the software solution extracts the customers for which the membership grade is between 0.4 and 0.6. This interval may be closer.

5. CONCLUSION

These paper theoretically and practically presented the integration knowledge discovery in databases, data mining and marketing research. The data mining component of the knowledge discovery process is mainly concerned with algorithms by which patterns are extracted from the data (fuzzy c-means clustering). Data for customers clustering are stored in relational data warehouse that is temporarily loading from transactional data bases. After running the program for fuzzy c-means clustering written in Visual Basic development environment follow the results that is easy understand and explain. Finally step is finding out suspicious records, customers for which the membership degree is close to 0.5. Expert judgment is necessary to analyze this tuples. On the business market the firm may very easy to define the required number of clusters (five, ten, twenty etc.) and the software will assign the customers to adequate cluster with membership degree. So has been assured the sensitivity and broad applicability the software and the concept of knowledge discovery.

Literature

2. G. Berks, D.G. Keyserlingk, J. Jantzen, M. Dotoli, H. Axer, Fuzzy Clustering- A Versatile Mean to Explore Medical Database, ESIT2000, Aachen, Germany