A Rough Set Approach towards Analysis of Cosmetic Data

P.M. Prasuna
Research Scholar
JNTU, Hyderabad

Y. Ramadevi,
Professor, Dept of CSE
CBIT, Hyderabad

A. Vinaya Babu,
Principal, JNTU, Hyderabad

ABSTRACT

The Global cosmetic industry borders are expanding everyday with new product launches and new packaging for old products. To cater the needs of the diverse customers, the industry has to deal with enormous features of skin, extracted from the face images of the customers, out of which few are required to identify the skin problems and necessary product rejuvenation. This paper proposes data mining techniques to meet these challenges in the cosmetic industry. The proposed approach is based on rough set theory, for discretizing the skin data which is of continuous valued in nature, feature selection using quick reducts and categorization of skin problems using rough clustering technique.

General Terms

Cosmetic industry, face image, continuous valued data, clustering.

Keywords

Rough set theory, discretization, feature selection, quick reducts, rough clustering.

1. INTRODUCTION

Bearing a long glowing heritage of cosmetic and beauty, aesthetic makeup products are being used since olden days. While the demand of beautifying substances are growing day by day, a large number of local as well as international manufacturers gradually extend their ranges and products in different provinces of the world. The product rejuvenation is based on intensive research results on skin analysis. The factors that lead to skin disorder are blood vessel diseases, diabetes, heart diseases, liver diseases, nutritional deficiencies, obesity, reactions to medications, stress, allergies to plants and other substances, climate, clothing, exposures to industrial and household chemicals, indoor heating. The main objective of the cosmetic industry is to discover the new product or renewal the existing product for a particular skin disorder. In order to do so they capture the face images of the their customers analyze the skin type find the root cause(s) for the disorders then release necessary products for reducing the identified disorder. In this research they have to handle enormous customer data with large number of features. Mining customer data with large number of features have some challenges. Firstly, the data is continuous valued which needs to be discretized [1] for further mining. All the features which are collected from customer face images may not be relevant. Finding irrelevant features is another challenge to meet. After feature selection the data is to be grouped basing on the feature which strongly influences the skin disorder. This provides the basis for categorization of the skin disorder. Basing on the skin disorder and influenced factors product innovation happens.

This paper proposes techniques based on Rough set theory for discretization, feature selection and clustering the customer skin data. This paper is organized as follows: in section 2 the Rough set Theory basics are outlined. Discretization techniques are presented in section 3, Feature selection techniques are discussed in section 4, Rough Clustering in section 5, the modeling process is detailed in section 6, and section 7 and 8 provide results and a brief conclusion is given.

2. ROUGH SET THEORY

Rough set theory is a mathematical method to handle imperfect data, proposed by Pawlak in 1982 [2]. It is best suited for the situation where the information available is insufficient to define a concept. The main objective of the rough set theory is to approximate a concept with respect to the information available. The basic concept of rough set theory is described from [2][3]. The two types of approximations used in rough set theory are lower approximation and upper approximation [2]. Lower approximation is the description of set of objects that certainly belong to the concept whereas the upper approximation describes the set of objects that possibly belong to the concept. The mathematical representation of Rough set is presented below [3].

An information system is a quadruple $S = \{U, A, V, f\}$ where U is the universe of finite set of objects, A is a non empty finite set of attributes, $A = \{a1, a2, a3, \dots, am\}$ Va is the domain of attribute a, $V = \cup_{a \in A} V_a$ and $f: U \times A \to V$ is a total function such that $f(x, a) \in$ Va, for each $a \in A, x \in U$(u,a) called knowledge or information function. Let R be a binary relation $R \subseteq U \times U$ and also an indiscernible relation. Let X is subset of U. To characterize X with respect to R its lower approximation ($R_*(X)$) and upper approximation ($R^*(X)$) are given as

$$R_*(X) = \{x : R(x) \subseteq X\}$$

$$R^*(X) = \{x : R(x) \cap X \neq \phi\}$$

The pair $(R_*(X), R^*(X))$ containing lower approximation and upper approximation is termed as a Rough set [4].

The boundary region of set X is defined by

$$RN_R(X) = R^*(X) - R_*(X)$$

The four classes of imprecision are defined as

1. Iff $R_*(X) \neq \emptyset\ and\ R^*(X) \neq U$ then X is roughly R definable
2. Iff $R_*(X) = \emptyset\ and\ R^*(X) \neq U$ then X is internally R undefinable.
3. Iff $R_*(X) \neq \emptyset\ and\ R^*(X) = U$ then X is externally R undefinable.
4. Iff $R_*(X) = \emptyset\ and\ R^*(X) = U$ then X is totally R undefinable.

The accuracy of the Rough Set is defined by

$$\alpha R(X) = \left|\frac{R_*(X)}{R^*(X)}\right|$$

The accuracy $\alpha R(X)$ lies between 0 and 1.If $\alpha R(X)$ = 1 then the set X is crisp otherwise it is Rough with respect to R.

3. DISCRETIZATION

In cosmetic industry the data collected is face images of the customers which are further processed to extract the features. These extracted features are mostly continuous valued. As it is evident that the continuous valued variables occupy large storage space and also leads to wrong interpretation and yields less precise clusters, it is essential that these variables to be discretized[5]. The discretization process has to generate reasonable number of cut points simultaneously maintaining the less information loss involved in the process of converting continuous valued data to discretized or nominal attributes[6].Rough set theory is an efficient approach to generate the cut points with less information loss[7]. The process is described below.

1. Take an information table I =(U,A) where U is the finite set of objects and A the set of attributes or features.
2. Any cut (a;c) on attribute a ∈ A divides V_a into two partitions and also creates two disjoint sets of U. Therefore a(U)=$\{v_1^a, v_2^a, v_3^a, v_4^a, \dots v_{na}^a\}$
3. We also consider that these values are arranged in sorting order i.e. $v_1^a < v_2^a < v_3^a < v_4^a, \dots < v_{na}^a$.
4. Now define Boolean variables defined by the information table I as $$BCuts_{I} = \{P_1^a, P_2^a, P_3^a, \dots P_n^a\}$$
5. Where P_k^a be a propositional variable corresponding to the interval $[v_k^a, v_{k+1}^a)$ for any $k \in \{1, \dots n_a - 1\}$ and $a \epsilon A$
6. Now create a new information table using the Boolean variables.
7. Then find the discernibility matrixes for the newly created table, from that generate discernibility formulae and prime implicants which generate the optimal cut points [8].

4. FEATURE SELECTION

4.1 Feature Selection using Reducts

Many times the data collected from customer face images is surplus. There would be inappropriate attributes or features without them the knowledge retrieval that is grouping people with similar skin disorders and also discovering skin disorder of the customers can be done easily. Feature selection is a process which ascertains the irrelevant features and eliminates them without disturbing the relevant features. This is a vital step in the analysis of skin type, as it aims to retain the discriminatory power of original features, thereby enhancing the accuracy and efficiency of discovering the skin behaviour of the customer. The residual set of attributes after feature selection is termed as reduct. Hence a reduct represents the minimal set of features[9]. There may be any number of reduct sets which would preserve the knowledge contained by the information system. Core is formed from the set of all reducts by considering only the common attributes. Further taking away any attribute from core will certainly leads to loss of knowledge of the information system. The basic concepts of rough set method useful for deriving a core are discussed below:

I. Let $S = \{U, A\}$ be an information system where U is a non empty finite set of object and A is a non empty finite set of attributes such that $a: U \rightarrow V_a$ for every $a \in A. V_a$ is the set of values that attribute a may take.

II. Let P be any subset of A. The associated equivalence relation is IND(P)[10].
$$IND(P) = \{(x, y) \in U^2 \,|\forall_a \in P, a(x) = a(x) = a(y)\}$$

III. The partition caused by $IND(P)$ is denoted by $U/IND(P)$ or U/P.

IV. Let 'a' is an attribute belongs to P .And it is said to be dispensable if $U/(P - \{a\}) = U/P$.

V. If all the attributes contained by set P are dispensable then P is said to be independent.

VI. Let P' be the subset of P. If P' is independent and $U/P = U/P'$ then P' is the reduct set of P.

VII. The intersection of all reducts is called the core of P i.e., Core (P) =∩Red (P)

```
FIND_REDUCT(U,P)
a) U, information table
b) P, the set of all features;
c) R ← P
d) Do
e) T ← R
f) T ← R-{x}
g) R ← T
h) Until $U/R = U/P$
i) return R
```

5. ROUGH CLUSTERING

The main task in cosmetic product development is to group the customers having similar disorders and analyze the features they are commonly possessing. This leads to the cosmetic product rejuvenation which helps in reducing the disorder the customers are suffering with. Clustering is the process of organizing the objects that have similar properties in groups[11]. Estimation of similarity among the objects is done basing on the attributes values that best describes the objects [12]. The rough set approach is combined with classic k means algorithm [13] to achieve more precise clusters.

Classic k-means algorithm:

1. Initially choose K points into the space represented by objects at random. These represent the “temporary” centroids of the Clusters.
2. Allocate each object to the cluster that has the closest centroid.
3. When all objects have been assigned then recalculate the centroid of each cluster as

$$x_j = \frac{\sum_{v \in X} v^j}{Size\ of\ cluster\ X}$$

Where $1 < j < m$

4. Reallocate the objects basing on the recalculated centroids.
5. Repeat the steps 3 & 4 until the centroids no longer move.

Applying Rough set method on k-Means: Here mainly the basic concepts of rough set theory the lower and approximations are incorporated[14]. The process of Rough k means follows the following procedure:

1. For each cluster the lower approximation and upper approximations are examined. If they are equal then the cluster is a conventional cluster. Centroid is calculated as

$$x_j = \frac{\sum v \in A_*(X)^{v^j}}{|A_*(X)|}$$

2. If both approximations are not equal, then the object has to be allocated to either lower approximation or upper approximation of the cluster.
3. Usually the object is assigned to lower approximation if the distance between the object and the cluster centroid smaller than the distance between object and any other cluster centroid.
4. To determine this first calculate the nearest centroid as

$$d_{min} = d(v, x_i) = \min_{1 \le j \le k} d(v, x^j)$$

If no other cluster is close to this we assign it to the lower approximation of the particular cluster, if any other cluster/clusters are nearer to it then we allocate the object to upper approximation of the corresponding clusters[15].

6. MODELING PROCESS

The analysis of cosmetic data involves different phases of collecting face images of customers, extracting features, discretization of numeric features, finding most relevant features and clustering[18] [19]. The modeling process is depicted as follows:

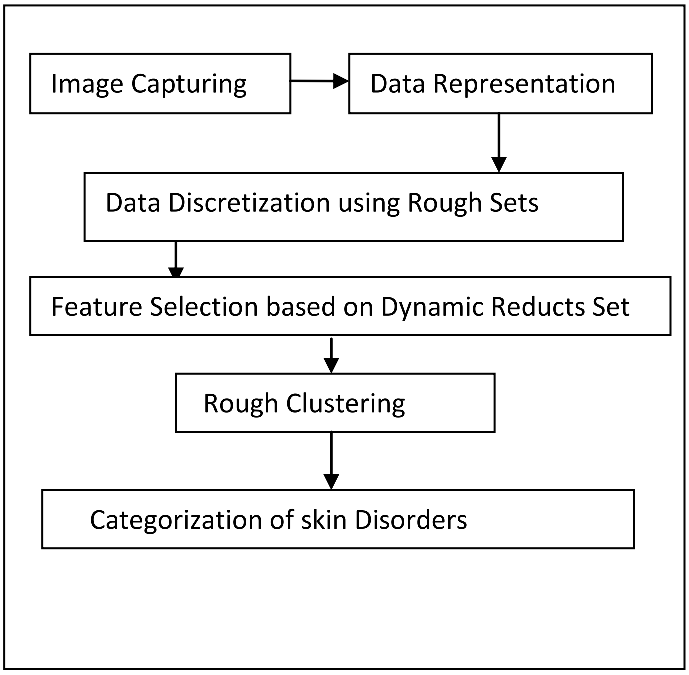


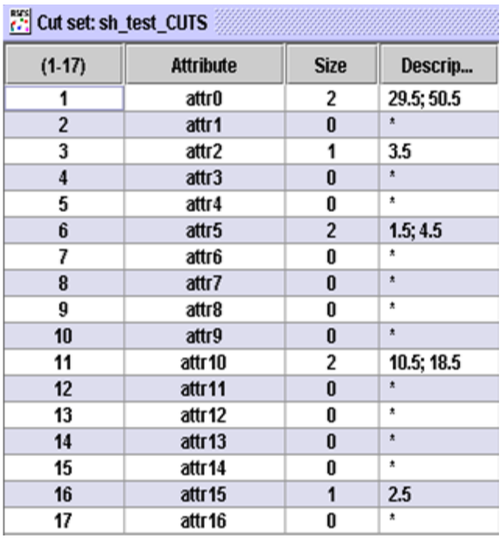
Image Capturing
Data Representation
Data Discretization using Rough Sets
Feature Selection based on Dynamic Reducts Set
Rough Clustering
Categorization of skin Disorders

Fig – 1 Modeling process of analysis of Cosmetic Data

7. EXPERIMENTAL RESULTS

In this section the results of the algorithms described in the previous section on cosmetic data are presented. A sample data set is given in TABLE -1.the algorithms are applied on the data set using RSES tool [16].

The discretization after applying rough set approach is given below:



Cut set: sh_test_CUTS

(1-17)	Attribute	Size	Descrip...
1	attr0	2	29.5; 50.5
2	attr1	0	*
3	attr2	1	3.5
4	attr3	0	*
5	attr4	0	*
6	attr5	2	1.5; 4.5
7	attr6	0	*
8	attr7	0	*
9	attr8	0	*
10	attr9	0	*
11	attr10	2	10.5; 18.5
12	attr11	0	*
13	attr12	0	*
14	attr13	0	*
15	attr14	0	*
16	attr15	1	2.5
17	attr16	0	*

Fig – 2 Discretization of the features of cosmetic generated using RSES Tool

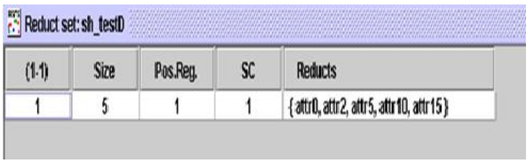
Table – 1 Features extracted from face images of customers in Cosmetic Industry
Sample Data Set

Image ID	age	Skin photo type	Spot Count	Age Spot	Pimples	Pastules	Papules	Area affected	Under Eye Wrinkles	Well hydrated skin	Cysts	Break Out Visibility	Acne count	pore count	Visible pores	Emerging lines	Fine Lines	Deep lines
1	22	5	59	3	4	0	0	1.5	0	63.2	0	3.96	12	1204	2.8717	44	16	5
20	23	5	52	2	6	0	0	14.3	0	77.2	0	1.55	8	1597	3.0032	32	9	6
23	25	6	49	1	5	0	1	0.8	0	80.4	0	1.89	8	1200	2.5261	57	10	1
24	22	5	21	1	0	0	0	1.1	0	75.4	0	0	10	1351	2.5792	42	7	2
29	22	5	16	1	0	0	0	5.7	0	71.7	0	0	1	867	1.8601	18	4	5
58	25	5	44	2	3	0	0	2.6	0	74.9	0	1.53	10	1197	2.6867	33	9	0
67	20	5	44	2	2	0	0	6.2	0	75.3	0	1.371	9	1225	2.5490	19	4	3
88	25	5	115	5	9	0	1	3.8	0	55.3	0	2.56	22	1512	3.1148	72	8	0
91	23	4	21	1	0	0	0	3.9	0	77.3	0	0	1	427	1.0588	38	11	3
33	29	1	5	1	0	0	0	24.6	0	84.7	0	0	0	746	1.4453	11	1	0
46	29	5	88	3	2	0	0	4.5	0	66.3	0	4.2	23	1468	3.1759	80	16	1
47	33	6	41	2	2	0	0	0.3	0	78.4	0	2.597	4	1372	3.4534	67	14	2
49	32	5	91	6	3	0	0	10.3	0	61.1	0	5.70	5	1263	2.6082	30	13	41
69	33	5	41	3	0	0	0	3.4	0	61.6	0	0	4	1135	2.3677	29	3	3
72	29	5	78	4	6	0	1	3.8	0	72.3	0	3.21	24	1170	2.546218	37	12	11
77	35	5	53	2	0	0	0	2.6	0	58.3	0	0	6	1140	2.3873	27	3	0
85	35	6	119	10	15	0	0	2.4	0	55.1	1	3.10	20	984	4.5490	99	57	18
95	35	5	89	4	1	0	0	0.4	0	65	0	1.53	24	1881	5.9193	81	23	12

Table – 2 Details of discretization of attributes

Attribute	Age	Redness	Pimples	Acne count	Deep lines
#of intervals	2	1	2	2	1
Interval value	[29.5, 50.5]	3.5	[1.5 4.5]	[10.5 18.5]	[2.5]

Results after applying quick reduct algorithm



Reduct set:sh_testD

(1-1)	Size	Pos.Reg.	SC	Reducts
1	5	1	1	{attr0, attr2, attr5, attr10, attr15}

Fig – 3 Reduct set generated in RSES Tool.

The reduct found from the 18 attribute set is
{Age, pimples, Marks from Acne count, Deep lines}

8. CONCLUSION

By examining the experimental results it is conclude that rough set approach is an effective and efficient way to discretize the continuous valued data and to remove the irrelevant features which enable a better mining process. This process can also be extended to distributed environment as the cosmetic industry is vigorously spanning its existence to various locations globally.

9. REFERENCES

[1] D Sotiris Kotsiantis, Dimitris Kanellopoulos "Discretization Techniques: A recent survey" GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 47-58

[2] Zdzisław Pawlak "Rough set theory for intelligent industrial applications " Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Poland

[3] Jan Komorowski, Lech Polkowski,Andrzej Skowron "Rough Sets: A Tutorial" Fundamenta Informaticae

[4] Zdzisław Pawlak "Rough set theory and its applications", Journal of Telecommunication and information Technology, 03/2002.

[5] Daniela Joiţa “Unsupervised static discretization methods in data mining” Titu Maiorescu University, Bucharest, Romania

[6] Girish Kumar Singh, Sonajharia Minz “Discretization Using Clustering and Rough Set Theory” Proceedings of the International Conference on Computing: Theory and Applications (ICCTA'07)

[7] Guan Xin, Yi Xiao, “Discretization of continuous interval-valued attributes in rough set theory and its application “ Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007

[8] Chengdong Wua, Mengxin Lia, Zhonghua Hana, Ying Zhanga Yong Yueb “Discretization Algorithms of Rough SetsUsing Clustering “Proceedings of the 2004 IEEE International Conference on Robotics and Biomimetics August 22 - 26, 2004, Shenyang, China

[9] Darshit Parmar, Teresa Wu , Jennifer Blackhurst “MMR: An algorithm for clustering categorical data using Rough Set Theory” Elsevier

[10] Andrzej Skowron1 and James F. Peters “Rough Sets: Trends and Challenges Extended Abstract Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing” Lecture Notes in Computer Science Volume 2639, 2003, pp 25-342 XXI (2001) 1001–1025 1001 IOS Press

[11] Shampa Sengupta1 and Asit Kumar Das “ Dimension Reduction Using Clustering Algorithm and Rough Set Theory” B.K. Panigrahi et al. (Eds.): SEMCCO 2012, LNCS 7677, pp. 705–712, 2012.© Springer-Verlag Berlin Heidelberg 2012

[12] E. Mohebi, M.N.N. Sap “ Rough set Based Clustering of the Self Organizing Map” 2009 First Asian Conference on Intelligent Information and Database Systems

[13] Approximate Distributed K-Means Clustering over a Peer-to-Peer Network IEEE transactions on knowledge and data engineering, vol. 21, no. 10, october 2009

[14] Tutut Herawan “Rough Clustering For Cancer Datasets” International Conference Mathematical and Computational Biology 2011 International Journal of Modern Physics: Conference Series Vol. 9 (2012) 240–258

[15] Chengdong Wua, Mengxin Lia, Zhonghua Hana, Ying Zhanga Yong Yueb “Discretization Algorithms of Rough Sets Using Clustering “ Proceedings of the 2004 IEEE International Conference on Robotics and Biomimetics August 22 - 26, 2004, Shenyang, China

[16] Jan G. Bazan1 and Marcin Szczuka2 “The Rough Set Exploration System” Transactions on Rough Sets III, LNCS 3400, pp. 37–56, 2005.Springer-Verlag Berlin Heidelberg 2005

[17] Zhengyou Zhou, Liusheng Huang, Ye Yun “Privacy Preserving Attribute Reduction Based on Rough Set “Second International Workshop on Knowledge Discovery and Data Mining 978-0-7695-3543-2/09

[18] Syed Sibte Raza Abidi, Kok Meng Hoe, Alwyn Goh “ Analyzing Data Clusters: A Rough Set Approach to Extract Cluster-Defining Symbolic Rules “ Fourth International Conference (IDA-01),2001, Cascais Portugal. Springer Verlag: Berlin.

[19] Roman W.Winiarski “ rough sets methods in feature reduction And classification” Int. J. Appl. Math. Comput. Sci., 2001, Vol.11, No.3