Association analysis of obesity/overweight and breast cancer using data mining techniques

Mahsa Dehghani Soufi¹, Reza Ferdousi^{1*}

¹Department of Health Information Technology, School of Management and Medical Informatics, Tabriz University of Medical Sciences, Tabriz, Iran

Article Info

Article type:
Research

Article History:
Received: 2020-10-19
Accepted: 2021-01-10
Published: 2021-02-14

*** Corresponding author:**
Reza Ferdousi

Department of Health Information
Technology, School of Management
and Medical Informatics, Tabriz
University of Medical Sciences,
Tabriz, Iran

Email: ferdousir@tbzmed.ac.ir

Keywords:
Breast Cancer
Overweight
Obesity
K-means
Apriori

ABSTRACT

Introduction: Growing evidence has shown that some overweight factors could be implicated in tumor genesis, higher recurrence and mortality. In addition, association of various overweight factors and breast cancer has not been extensively explored. The goal of this research was to explore and evaluate the association of various overweight/obesity factors and breast cancer, based on obesity breast cancer data set.

Material and Methods: Several studies show that a significantly stronger association is obvious between overweight and higher breast cancer incidence, but the role of some overweight factors such as BMI, insulin-resistance, Homeostasis Model Assessment (HOMA), Leptin, adiponectin, glucose and MCP.1 is still debatable. So for experiment of research work several clinical and biochemical overweight factors, including age, Body Mass Index (BMI), Glucose, Insulin, Homeostatic Model Assessment (HOMA), Leptin, Adiponectin, Resistin and Monocyte chemo attractant protein-1 (MCP-1) were analyzed. Data mining algorithms including k-means, Apriori, Hierarchical clustering algorithm (HCM) were applied using orange version 3.22 as an open source data mining tool.

Results: The Apriori algorithm generated a list of frequent item sets and some strong rules from dataset and found that insulin, HOMA and leptin are two items often simultaneously were seen for BC patients that leads to cancer progression. K-means algorithm applied and it divided samples on three clusters and its results showed that the pair of <Adiponectin, MCP.1> has the highest effect on separation of clusters. In addition HCM was carried out and classified BC patients into 1-32 clusters to So this research apply HCM algorithm. We carried out hierarchical clustering with average linkage without punning and classified BC patients into 1-32 clusters in order to identify BC patients with similar characteristics.

Conclusion: These finding provide the employed algorithms in this study can be helpful to our aim.

Cite this paper as:

Dehghani Soufi M, Ferdousi R. Association Analysis of Obesity/Overweight and Breast Cancer Using Data Mining Techniques. Front Health Inform. 2021; 10: 60. DOI: [10.30699/fhi.v10i1.255](https://doi.org/10.30699/fhi.v10i1.255)

INTRODUCTION

In the developed world, the majority of the adults population falls in the overweight and obese categories as determined by body mass index (BMI: 25-29.9 and >30, respectively) [1]. Unfortunately, some factors such as nutrition, economic and changes in technologies increasing the obesity prevalence. Some malignances are directly related to obesity, such as breast cancer. Breast cancer is the leading cause of cancer-related mortality worldwide [2, 3]. However, the American instate for cancer research

highlighted that being overweight/obese decreased the risk of breast cancer [4]. The overweight factors and breast cancer have been the subject of many studies but the role of some clinical and biochemical factors such as BMI, insulin-resistance, Homeostasis Model Assessment (HOMA), Leptin, adiponectin, glucose and MCP.1 is still debatable [5]. Growing evidence has shown that these alterations could be implicated in tumor genesis, higher recurrence and mortality [6, 7].

In recent years, data mining techniques has been

applied to solve various problems especially in health care. Data mining algorithms has been greatly improved in order to make support doctors making decision by analyzing data and discovering patterns in present datasets [8]. Analyzing the obesity factors in order to discover new patterns can be helpful for clinicians. So we hypothesize that the breast cancer distribution in overweight and obese women may differ from normal weight and this puts them at higher risk for breast cancer. Furthermore, association of above mentioned various overweight factors and breast cancer has not been extensively explored. In this study, we intended to determine the relation between these characteristics in a group of overweight/obese women breast cancer.

MATERIAL AND METHODS

For the experimental purpose we use an obesity breast cancer project data file to analyze, this dataset that contains 64, include both continuous and

categorical features and it was clear and there are not any missing values and duplicate data. To perform experimental works of this study, take orange version 3.22 as an open source data mining tool and then apply data mining algorithm and different evaluation methods. We aim to identify groups of obese BC patients with similar characteristics, clusters by using a clustering method and make a comparison between them. We have use data preprocessing method and the method that we were going to handle is K-means. In addition two different association mining algorithm was used in this work. The data set attributes descriptions are as below in Table 1.

Preprocessed data is shown in Fig 1. A complete overview of data mining process which was followed in current project, is displayed in schematic model (Fig 2). We have described these pathways in the methods section separately.

Table 1: the obese breast cancer data set attribute

Attribute name	Description
Age	Patients' age in years
BMI	Body Mass Index is a measure of body fat based on height and weight that applies to adult men and women $m = \text{mass (in kilograms)}$ $h = \text{height (in meters)}$ $BMI = \frac{m}{h^2}$
Glucose	A carbohydrate compound that get from foods and body uses it for energy
Insulin	Hormone that allows your body to use sugar (glucose)
HOMA	Homeostasis Model Assessment (HOMA) index was calculated to evaluate insulin resistance
Leptin	hormones directly connected to body fat and obesity and released from the fat cells located in adipose tissues
Adiponectin	A protein hormone produced and secreted exclusively by adipocytes (fat cells) that regulates the metabolism of lipids and glucose.
Resistin	or 'resistance to insulin: is a cysteine-rich peptide hormone causes resistance of peripheral tissues to insulin
MCP.1	The monocyte chemoattractant protein-1 has role in fat storage, adipocytes synthesize.

Data instances: 64
 Features: 9
 Meta attributes: None

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
1	45	21.3039	102	13.852	3.48516	7.6476	21.0566	23.03408	552.444
2	45	20.83	74	4.560	0.832352	7.7529	8.23741	28.03230	382.955
3	49	20.9566	94	12.305	2.85312	11.2406	8.41217	23.11770	573.630
4	34	24.2424	92	21.699	4.92423	16.7353	21.8237	12.06534	481.949
5	42	21.3599	93	2.999	0.687971	19.0826	8.46292	17.37615	321.919
6	68	21.0828	102	6.200	1.55992	9.6994	8.57465	13.74244	448.799
7	51	19.1327	93	4.364	1.0011	11.0816	5.80762	5.57055	90.600
8	62	22.6562	92	3.482	0.790182	9.8648	11.2362	10.69548	703.973
9	38	22.4996	95	5.261	1.23283	8.4380	4.77192	15.73606	199.055
10	69	21.5139	112	6.683	1.84629	32.5800	4.13802	15.69876	713.239
11	49	21.3675	78	2.640	0.507936	6.3339	3.88614	22.94254	737.672
12	51	22.8928	103	2.740	0.696143	8.0163	9.34977	11.55492	359.232
13	59	22.8329	98	6.862	1.65877	14.9037	4.23011	8.20490	355.310
14	45	23.1405	116	4.902	1.40263	17.9973	4.29471	5.26330	518.586
15	54	24.2188	86	3.730	0.791257	8.6874	3.70523	10.34455	635.049
16	64	22.2222	98	5.700	1.37788	12.1905	4.78399	13.91245	395.976
17	46	20.83	88	3.420	0.742368	12.8700	18.55	13.56000	301.210
18	44	19.56	114	15.890	4.46827	13.0800	20.37	4.62000	220.660

Fig 1: A screenshot of selected obese breast cancer dataset

Fig 2: Tree pathways of current projects

In preprocessing step of dataset, considering the total number of features (F) in dataset, we have discretized features in different groups by using Equal-width discretization method as displayed in Fig 3.

Age: 44.40, 54.80, 65.20, 75.60 (equal width k=5)
 BMI: 22.12, 25.87, 29.61, 33.36 (equal width k=5)
 Glucose: 96.20, 122.40, 148.60, 174.80 (equal width k=5)
 Insulin: 16.44, 30.45, 44.45
 HOMA: 6.64, 12.78, 18.91
 Leptin: 27.32, 48.31, 69.29
 Adiponectin: 9.68, 17.70, 25.73
 Resistin: 16.21, 29.21, 42.21
 MCP:1: 492.18, 894.27, 1296.35

Fig 3: Discretization results

There are 64 patients in this dataset and the result of K-means clustering and association rule mining methods gives separated pathway after preprocessing (Fig 2). Fig 4 shows the working model of this research work.

Fig 4: Data mining process model in current research

RESULTS

First pathway: Association Rule Mining implementation

ARM is a method to find the associations and/or relationships among items in large databases. So, we can use it to detect relations among inputs of any system and later eliminate some unnecessary inputs.

The Apriori algorithm is a state of the art algorithm most of the association rule algorithms are somewhat variations of this algorithm. The Apriori algorithm works iteratively. It first finds the set of large 1-item sets, and then set of 2- item sets, and so on. The number of scan over the transaction database is as many as the length of the maximal item set. Apriori is based on the following fact: The simple but powerful observation leads to the generation of a smaller candidate set using the set of large item sets found in the previous iteration [9]. The Apriori algorithm presented as follows:

```

Apriori()
L1 = {large 1-itemsets}
k = 2
while Lk-1 ≠ ∅ do
  begin
    Ck = apriori_gen(Lk-1)
    for all transactions t in D do
      begin
        C' = subset(Ck, t)
        for all candidate c ∈ C' do
          c.count = c.count + 1
        end
      end
    Lk = {c ∈ Ck | c.count ≥ minsup}
    k = k + 1
  end
end
  
```

So, we applied Apriori algorithm in form of a pre-defined library of Orange Canvas. The algorithm generated a list of frequent item sets and some strong rules from dataset. The strongest rule was:

Insulin ≤ 16.439 → HOMA ≤ 6.64354

It was described that every patient with an insulin value of less than 16.439, has a HOMA value less than 6.64 with support and confidence 75% and 100%, respectively. The results were shown in Table 3 and 4. Table 2 showed rules with minimum support value of .30.

In Table 2 and figure 5, we listed some meaningful and valuable association rules found between overweight factors in BC patients. For example, in this study we found that in 48 samples of total with support=75% and confidence=100% every patient with Insulin ≤ 16.439 has HOMA ≤ 6.64. This means insulin and HOMA are two items often simultaneously were seen for BC patients and we find out insulin, HOMA and leptin with support=60% and confidence= 95% are item sets related to breast cancer development and progression. We can report a positive association between obesity (patient with pre-diabetes) and cancer, regardless of these factors, diabetes may be an independent risk factor for cancer.

This proves the method implemented in current study can indeed find meaningful and valuable association rules to provide useful suggestions for relevant treatment.

Second pathway: Implementation using K-means method

Steps of k-means algorithm was including:

Input: Input breast cancer dataset

The first step of K-means clustering is to define the number of the cluster and their centroids (In this case k=3).

K clusters are created by associating every data point with the nearest mean.

The centroid becomes the new means

Steps (b) and (c) are repeated until convergence has

been reached [10]. Frequent item sets of cancer diagnosed patients is presented in Fig 5. Results obtained applying this algorithm is shown in Fig 6.

The results showed that the pair of <Adiponectin, MCP.1> has the highest effect on separation of clusters.

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent	Consequent
0.750	1.000	0.750	1.146	1.164	0.105	Insulin=< 16.439	HOMA=< 6.64354
0.609	0.951	0.641	1.341	1.107	0.059	Leptin=< 27.3204	HOMA=< 6.64354
0.578	0.902	0.641	1.171	1.203	0.098	Leptin=< 27.3204	Insulin=< 16.439
0.578	1.000	0.578	1.486	1.164	0.081	Insulin=< 16.439, Leptin=< 27.3204	HOMA=< 6.64354
0.578	0.902	0.641	1.171	1.203	0.098	Leptin=< 27.3204	Insulin=< 16.439, HOMA=< 6.64354
0.578	0.949	0.609	1.231	1.265	0.121	HOMA=< 6.64354, Leptin=< 27.3204	Insulin=< 16.439
0.469	1.000	0.469	1.833	1.164	0.066	Insulin=< 16.439, Resistin=< 16.21132	HOMA=< 6.64354
0.469	0.909	0.516	1.455	1.212	0.082	HOMA=< 6.64354, Resistin=< 16.21132	Insulin=< 16.439
0.469	1.000	0.469	1.833	1.164	0.066	Insulin=< 16.439, Adiponectin=< 9.67952	HOMA=< 6.64354
0.469	0.938	0.500	1.500	1.250	0.094	HOMA=< 6.64354, Adiponectin=< 9.67952	Insulin=< 16.439
0.422	1.000	0.422	2.037	1.164	0.059	Insulin=< 16.439, MCP.1=< 492.178	HOMA=< 6.64354
0.422	0.900	0.469	1.600	1.200	0.070	HOMA=< 6.64354, MCP.1=< 492.178	Insulin=< 16.439
0.391	0.926	0.422	2.037	1.077	0.028	Leptin=< 27.3204, Resistin=< 16.21132	HOMA=< 6.64354
0.391	0.926	0.422	2.037	1.077	0.028	Glucose=96 - 122	HOMA=< 6.64354
0.391	0.962	0.406	2.115	1.119	0.042	Glucose=< 96	HOMA=< 6.64354
0.375	1.000	0.375	2.000	1.333	0.094	Leptin=< 27.3204, Adiponectin=< 9.67952	Insulin=< 16.439
0.375	1.000	0.375	2.292	1.164	0.053	Leptin=< 27.3204, Adiponectin=< 9.67952	HOMA=< 6.64354
0.375	1.000	0.375	2.292	1.164	0.053	Insulin=< 16.439, Leptin=< 27.3204, Adiponectin=< 9.67952	HOMA=< 6.64354
0.375	1.000	0.375	2.000	1.333	0.094	Leptin=< 27.3204, Adiponectin=< 9.67952	Insulin=< 16.439, HOMA=< 6.64354
0.375	1.000	0.375	2.000	1.333	0.094	HOMA=< 6.64354, Leptin=< 27.3204, Adiponectin=< 9.67952	Insulin=< 16.439
0.359	1.000	0.359	2.391	1.164	0.051	Insulin=< 16.439, Leptin=< 27.3204, Resistin=< 16.21132	HOMA=< 6.64354
0.359	0.920	0.391	1.920	1.227	0.066	HOMA=< 6.64354, Leptin=< 27.3204, Resistin=< 16.21132	Insulin=< 16.439
0.344	0.917	0.375	2.292	1.067	0.021	Leptin=< 27.3204, MCP.1=< 492.178	HOMA=< 6.64354
0.344	1.000	0.344	2.500	1.164	0.048	Glucose=96 - 122, Insulin=< 16.439	HOMA=< 6.64354
0.344	1.000	0.344	2.500	1.164	0.048	Glucose=< 96, Insulin=< 16.439	HOMA=< 6.64354
0.328	1.000	0.328	2.619	1.164	0.046	Insulin=< 16.439, Leptin=< 27.3204, MCP.1=< 492.178	HOMA=< 6.64354
0.328	0.955	0.344	2.182	1.273	0.070	HOMA=< 6.64354, Leptin=< 27.3204, MCP.1=< 492.178	Insulin=< 16.439
0.312	1.000	0.312	2.750	1.164	0.044	Insulin=< 16.439, Resistin=< 16.21132, MCP.1=< 492.178	HOMA=< 6.64354
0.312	0.909	0.344	2.182	1.212	0.055	HOMA=< 6.64354, Resistin=< 16.21132, MCP.1=< 492.178	Insulin=< 16.439

Table 2: Association mining rules

Itemsets	Support	%
▲ Insulin=< 16.439	48	75
▲ HOMA=< 6.64354	48	75
▲ Leptin=< 27.3204	37	57.81
▲ Leptin=< 27.3204	37	57.81
▲ HOMA=< 6.64354	55	85.94
▲ Leptin=< 27.3204	39	60.94
▲ Resistin=< 16.21132	33	51.56
▲ Leptin=< 27.3204	41	64.06
▲ Adiponectin=< 9.67952	36	56.25
▲ Resistin=< 16.21132	39	60.94
▲ MCP.1=< 492.178	34	53.12

Fig 5: Frequent item sets

Fig 6: K-means results with a view on highest feature projection. It divided samples on three clusters.

Third pathway: Implementation using HCM (Hierarchical Clustered Method)

This research work used HCM algorithm for screening of obese breast cancer patients using orange environment. In this algorithm clusters are

created from previously initiated clusters. So this research apply HCM algorithm. We carried out hierarchical clustering with average linkage withouturning and classified BC patients into 1–32 clusters. By applying this algorithm the following outcome generated (Fig 7).

Fig 7: HCM clustering outcome

DISCUSSION

In previous years, a lot of similar works has been done using data mining techniques in healthcare sector, the result of this work demonstrate the ability of applied algorithms that confirm the results those reported in [5], as the original creators of the used dataset, obtaining 95% confidence interval for sensitivity, specificity and AUC was (82%, 88%), (84%, 90%) and (0.87, 0.91) respectively (using SVM), also by Alickovic and Subasi [11] who used the Random Forest and Genetic Algorithm, obtaining the highest accuracy of 99.48 to diagnosis of breast

cancer. In addition the authors in [12] applied different classification model to find the best biomarker for breast cancer. The result of artificial neural network comparatively better than the KNN algorithms results and get an accuracy of 80%.

Some data mining techniques including Apriori, K-means and HCM algorithms were applied to illustrate how we can using these algorithms on datasets. Identifying which overweight factors robustly linked to breast cancer is critical for designing effective interventions, so good results obtained for our purpose too, and we believe that, a strict comparison could not be made between results obtained in other

studies, because the algorithms and aim of studies are different.

Table 3: Literature Review Summary Table

Authors	DM techniques	Research area	outcomes	Publication year
Patrício et al [5]	SVM	using obesity factors to predict the presence of breast cancer	Sensitivity: [82%, 88%] Specificity: [84%, 90%] AUC: [0.87, 0.91]	2018
Alickovic et al [11]	Random Forest and Genetic Algorithm	diagnosis of Breast cancer	Accuracy: 99.48%	2017
Ghani et al [12]	Classification techniques such as artificial neural network and KNN algorithms	find the best biomarker for breast cancer	Accuracy: 80 %	2019

The attractiveness of these algorithms is the ability to involve them in the health sector in order to support the decision making of clinicians, in this way, the process of diagnosis and treatment can be carried out with reliable results.

CONCLUSION

In this study we applied the data mining algorithms to evaluate the overweight/obesity factors on breast cancer data set. The outcome of the research is justified that k-means and Apriori algorithms are helpful to make a comparison between overweight factors in BC patients to provide useful suggestions for relevant treatment. In addition HCM algorithm was convenient for this research work to identify groups of overweight breast cancer patients with similar characteristics. The results of the original papers [5, 13] of this data set is consistent with that found on the present study. This issue should be considered in the design of future studies to confirm the presented results. Thus, our data need to be analyzed further and will require a larger sample in order to prove causality and provide further insights into the role of such parameters of obesity in breast malignancy.

AUTHOR'S CONTRIBUTION

The authors agree on this final form of the manuscript, and attested that all authors contributed in the final draft of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this study.

FINANCIAL DISCLOSURE

No financial interests related to the material of this manuscript have been declared.

REFERENCES

- Argolo DF, Hudis CA, Iyengar NM. The impact of obesity on breast cancer. *Current Oncology Reports*. 2018; 20(6): 47.
- Kang C, LeRoith D, Gallagher EJ. Diabetes, obesity, and breast cancer. *Endocrinology*. 2018; 159(11): 3801-12. PMID: 30215698 DOI: 10.1210/en.2018-00574 [PubMed]
- Torre LA, Islami F, Siegel RL, Ward EM, Jemal A. Global cancer in women: Burden and trends. *Cancer Epidemiol Biomarkers Prev*. 2017; 26(4): 444-57. PMID: 28223433 DOI: 10.1158/1055-9965.EPI-16-0858 [PubMed]
- American Institute for Cancer Research. Breast cancer [Internet]. 2015 [cited: 17 Aug 2019; updated: 9 Jan 2020]. Available from: <http://www.aicr.org/continuous-update-project/breast-cancer.html>.
- Patrício M, Pereira J, Crisóstomo J, Matafome P, Gomes M, Seica R, et al. Using resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer*. 2018; 18(1): 29. PMID: 29301500 DOI: 10.1186/s12885-017-3877-1 [PubMed]
- Cohen DH, LeRoith D. Obesity, type 2 diabetes, and cancer: The insulin and IGF connection. *Endocr Relat Cancer*. 2012; 19(5): F27-45. PMID: 22593429 DOI: 10.1530/ERC-11-0374 [PubMed]
- Crisostomo J, Matafome P, Santos-Silva D, Gomes AL, Gomes M, Patrício M, et al. Hyperresistinemia and metabolic dysregulation: A risky crosstalk in obese breast cancer. *Endocrine*. 2016; 53(2): 433-42. PMID: 26892376 DOI: 10.1007/s12020-016-0893-x [PubMed]
- Hossain R, Mahmud SH, Hossain MA, Noori SR, Jahan H. PRMT: Predicting risk factor of obesity among middle-aged people using data mining techniques. *Procedia Computer Science*. 2018; 132: 1068-76.
- Yazgana P, Kusakci AO. A literature survey on association rule mining algorithms. *Southeast Europe Journal of Soft Computing*. 2016; 5(1): 5-14.
- Dey A. Machine learning algorithms: A review. *International Journal of Computer Science and Information Technologies*. 2016; 7(3): 1174-9.

-
11. Aličković E, Subasi A. Breast cancer diagnosis using GA feature selection and Rotation Forest. *Neural Computing and Applications*. 2017; 28(4): 753-63.
 12. Ghani MU, Alam TM, Jaskani FH. Comparison of classification models for early prediction of breast cancer. *International Conference on Innovative Computing*. IEEE; 2019.
 13. Dalamaga M. Obesity, insulin resistance, adipocytokines and breast cancer: New biomarkers and attractive therapeutic targets. *World J Exp Med*. 2013; 3(3): 34-42. PMID: 24520544 DOI: 10.5493/wjem.v3.i3.34 [[PubMed](#)]