IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 8, Issue 5 (Jan. - Feb. 2013), PP 17-21
www.iosrjournals.org

SKM-A Conspicuous Way to Predict Frequent Item Sets

[1]A.Bamini, [2]Dr. S. Franklin John, [3]Dr. P. Ranjit Jeba Thangiah,
[1]*Assistant Professor, the Standard Fireworks Rajaratnam College for women, Sivakasi*
[2]*Professor and Principal, Nehru College of Management, Coimbatore*
[3]*Assistant Professor (SG), Karunya University, Coimbatore*

Abstract: *Market Basket Analysis is the general name for understanding product purchase patterns at the customer level in the super market. This paper focuses on clustering similar and frequent item sets using an integrated approach of SOM and K-Means clustering techniques. The proposed algorithm finds the number of clusters using Artificial Neural Network (SOM) and then groups the similar item sets into the respective clusters using the k-means to improve the marketing campaign.*
Keywords: *Clustering; k-means; SOM*

I. Introduction

Data Mining is the basic process employed to analyze patterns in data and extract information [1]. Data mining is actually the core of a larger process, known as knowledge discovery in databases (KDD). KDD is the process of taking low-level data and turning it into another form that is more useful, such as a summarization or a model. [2]

According to Insightful Miner 3.0 User Guide, "Data mining is the application of statistics in the form of exploratory data analysis and predictive models to reveal patterns and trends in very large data sets."According to Gartner Group, "Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques."

According to Hand "data mining is the process of secondary analysis of large databases aimed at finding unsuspected relationships which are of interest or value to the database owners." Data Mining has been referred as a statistical process of analyzing data stored in a warehouse (Decker, 1998).

As we know the importance of data mining in marketing area of banking industry, similar applications of data mining can also seen in marketing area of retail industry. In this "market basket analysis" is a marketing method used by many retailers to determine the optimal locations for the establishment of marts or shopping malls and promote their products. The basic idea behind this technique is to identify what type of products customers most often purchase together so that the promotion campaigns and advertisements are planned in such a way that maximum sales can be generated, thus increase the revenue of the firm.

Data mining tools proved to be the best way of discovering patterns in the data for the successful customer relations, since, nowadays business is relationship based the data is extremely important because retailers reach customers best through the data. Retailers can study customers' past purchasing histories and know with what kinds of promotions and incentives to target customers. By collecting customer and transactional data, data mining tools helps the retailers to identify best customers and offer exclusive extras and incentives to them. Data mining applications like Target marketing, which identifies the prospective customers by observing the past purchase behavior, can reduce the expenditure in campaigning and promotional activities; Attrition prediction and churn analysis used to prevent loss of customers and avoids adding churn-prone customers, this data mining application uses the neural nets and time series analysis, the ideal benefits includes retention of customers and more effective promotions.

II. Review Literature:

In order to use standard statistics, a technique would be needed that can handle both continuous and categorical variables and will create a model that will allow the classification of a new observation. According to Sharma (1996) [3], logistic regression would be the technique to use. In this, the best combination of variables is discovered that maximizes the correct predictions for the current set and is used to predict membership of the new observation. This methodology looks for the best combination of variables to produce a prediction. For this project, however, there will be different types of Web pages that are deemed appropriate, and thus it may prove difficult to converge on a single solution using logistic regression.

Memory-based reasoning is where a memory of past situations is used directly to classify a new observation. N-neighbor non-parametric discriminant analysis is one statistical technique used for MBR. This concept was discussed in 1988 by Stanfill and Waltz in The Memory Based Reasoning Paradigm at a DARPA

workshop. In MBR, some type of distance function is applied to judge the distance between a new observation and each existing observation, with optional variable weighting. The program then looks at a number of the preclassified neighbors closest to the new observation and makes a decision [4].

Neural networks are based on the workings of neurons in the brain, where a neuron takes in input from various sources, processes it, and passes it on to one or more other neurons. The neuron accepts 0-1 measurements of each variable. It then creates a hidden layer of neurons, which weights and combines the variables in various ways. Each neuron is then fed into an output neuron, and the weights and combinations of the neurons are adjusted with each observation in the training set through back-propagation until an optimal combination of weights is found [5].

The Neural networks are very versatile, as they do not look for one optimal combination of variables; instead, several different combinations of variables can produce the same result. They can be used in very complicated domains where rules are not easily discovered. Because of its ability to handle complicated problems, a neural network may be the best choice for this problem [4].

Plunkett's Four Keys to Successful Retailing:

- *A High Value-High Quality Product Selection:* Depth of selection is less important than a reasonably sized offering of products that the merchandiser has chosen because they consistently offer high value and quality.
- *Very Competitive Prices:* The goal here is to give the consumer confidence that the store faithfully delivers everyday low prices—meanwhile, managing the firm so as to allow the owners a viable profit margin.
- *Superior Service:* In-store help, follow up service, problem-solving, installation and repairs offered easily and quickly—the ability to make returns and exchanges must be part of the package, with an absolute minimum of inconvenience to the consumer.
- *Seamless Integration of Bricks and Clicks:* Successful firms integrate their online endeavors with their physical presence in a manner that provides the highest possible level of convenience to customers.

III. Related Works:

Self-organizing maps (SOMs) are a type of artificial neural network that functions as an unsupervised classifier with topology preservation. This means that a SOM creates a network that automatically arranges data based on its topology, an approach which can be useful for a variety of tasks including clustering, classification, non-linear PCA (Villmann, Wieland, & Michael, 2000), vector quantization (de Bodt, Verleysen, & Cottrell, 1997), and K-means clustering (Bacao, Lobo, & Painho, 2005)[6]. There are several variants of SOMs including some that dynamically increase the number of units and others that dynamically increase the spaces between units. Despite the diverse uses and variants of SOMs, they always contain units with values, a matching function, a selection function, and a value changing function. Unlike other methods that perform similar tasks, SOMs are highly scalable to large high-dimensional datasets. The size of the SOM can be scaled in relation to the dataset and desired results, with larger SOMs for finding differences and smaller SOMs for finding similarities.

Macqueen [7] developed the K-means algorithm that assigns each subject to its cluster center or mean. K-means clustering uses an iterative algorithm that minimizes the sum of distances from each subject to its cluster center over all clusters. It is often called the k-means model, since the cluster centers are the means of the subjects assigned to each cluster when the algorithm is run to complete convergence. We can control the minimization by using several optional input parameters, including the initial values of the cluster centers and the maximum number of iterations.

IV. Proposed Works

Preliminaries

SOM:

The Kohonen map is considered to be the original or traditional SOM and is composed of a two-dimensional grid of units. The grid can be composed of square or hexagonal units that are initialized using random numbers or eigen values from the data.

During a training process, the SOM is exposed to vectors from the training data one at a time and the SOM is adjusted to have more similar values. The training process begins by activating the single unit in the SOM that has the smallest Euclidean distance from an input vector. The active unit then creates a neighborhood by selecting all the adjacent units up to a certain distance. All the units in the neighborhood are then adjusted so that all their Euclidean distances from the input vector are smaller. Through repetition of this process the SOM forms and the data topology is imprinted [8].

The creation of a Kohonen map requires the user to specify the topology, neighborhood type, x-dimension, y-dimension, and the dimensionality of the SOM. The topology is the connections between the grid units visible at the adjacent edge and is usually rectangular or hexagonal. The neighborhood type is the way in

which connections are made between units and is usually Gaussian. The x-dimension and the y dimension are the number of units in the x-direction and y-direction. The dimensionality of the SOM is the number of variables the SOM is designed for, which depends exclusively on the number of variables in the data.

K-MEANS:

K-means [9] is also considered to be one of the important tools for clustering problems. K-means works using the following steps:

1. Place K objects points into the space that are to be clustered object points always represent initial group centroids.
2. Assign each object point to the group that has the closest centroid.
3. When all object points have been assigned, re-calculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the object points into groups from which the metric to be minimized can be calculated.

It is an indicator of the distance of the n data points from their respective cluster centers.

SKM Algorithm

This paper focuses on clustering similar transactions (item sets) and the diagrammatic representation is:

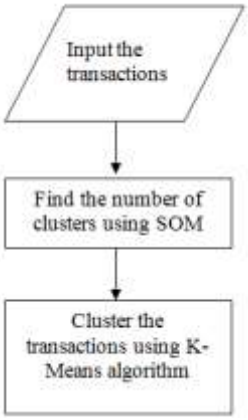

Fig-1: SKM

Step 1: Transaction data from a super market has been collected for nearly 2 months.

Step 2: After Data cleaning has been performed manually using Office package, around 1000 transactions has been identified. The essential dataset S_i has been segregated into Groceries, Cosmetics, Stationeries, Snacks and Household.

Step 3: Normalization has been carried out by converting categorical data to binary data.

Step 4: Kohonen's SOM algorithm has been used to find the number of clusters.

Step 5: The outcome of SOM is passed as input to k in K-Means algorithm along with the transaction dataset.

Step 6: Similar transactions are grouped together using k-means. The working of K-Means has been evaluated using Euclidean and Manhattan distance measures.

$$p_{ij=(\Sigma} |x_{ik} - x_{jk}|^{r}{}_{)}{}^{1/r}$$

where, *r* is a parameter, *xik* and *xjk* are data objects. The following is a list of the common Minkowski distances for specific values of *r*.

1) *r* = 1. City block (Manhattan, taxicab, L1 norm) distance.
2) *r* = 2. Euclidean distance.

Step 7: The k-Means technique minimizes the intra cluster variance, or squared error:

$$E = \Sigma \Sigma \|p - m_i\|^2$$

where there are k clusters Ci, i = 1,2, ...,k and mi is the centroid of all the points p ε Si.

Step 8: Negligible results are considered to be Outliers and are removed.

V. Findings And Results

The result has been obtained by developing SKM and has been checked by giving Groceries, Cosmetics, Household, Stationaries and Snacks real datasets collected from the super market.

Table 1: Results obtained from 5 datasets

DataSet	No. of clusters	Sample Frequent transactions (item sets) from clusters
Groceries	33	{ Rice Pappad Masala Powder Wheat Flour Raagi} { Cashew Cardamom Jaggery **}** **{** Urad dal Toor dal Tamarind Kala Channa Masala Powder Dry Chilli Fried Gram Anise Cumin Seeds Pepper Pappad Green Gram dal Mustard Seeds Oil Fenugreek Asafoetidia }
Cosmetics	6	{ Shampoo Oil Soap } { Cream Soap }
Household	17	{ Detergent Cake Detergent Powder } { Brush Paste Detergent Cake Dish Washer } { Paste Detergent Powder Dish Washer }
Stationeries	3	{ Pen Pepncil Paper Rubber Scale Sharpener Box Clip } { Pen Box Safety Pin Hair Pin Scissors Blade Cello Tape }
Snacks	15	{ Chocolate Biscuit Noodles Popcorn Burfi} { Biscuit Dates Noodles Popcorn Health Drinks Burfi Cool Drinks }

Frequent transactions are found to be more in Groceries rather than other datasets.

Table 2: Performance Table

Datasets	Metrics	KM-E	KM-M	SKM-E	SKM-M
Groceries	**Time**	6.141	6.116	4.900	4.883
	Error	6.3058	6.2996	4.0589	1.0343
Cosmetics	**Time**	0.924	0.912	0.885	0.861
	Error	4.3564	4.3223	2.2876	2.2665
Household	**Time**	1.578	1.549	1.323	1.320
	Error	3.9688	3.8056	2.8643	2.8428
Stationeries	**Time**	0.164	0.159	0.143	0.142
	Error	3.3393	3.1393	1.9318	1.8413
Snacks	**Time**	1.895	1.874	1.561	1.544
	Error	3.7126	3.6989	1.8905	1.8795

Table-2 shows the performance measure obtained using 5 datasets. In the table, KM-E represents K-means with Euclidean measure, KM-M represents K-means with Manhattan measure, the result obtained in this process is represented in SKM-E(SOM based K-means with Euclidean measure) and SKM-M(SOM based K-means with Manhattan measure).

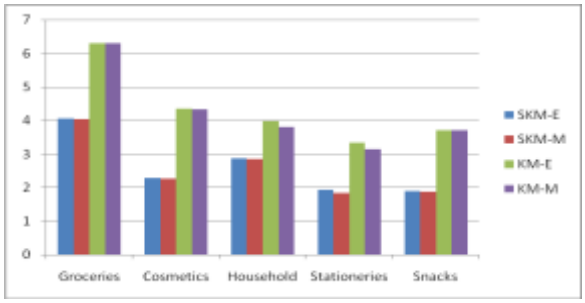

Fig-2: Average Error Parameter

Fig 2 and Fig 3 represents the graphical chart for the Table-2. Here the KM-E shows K-means with Euclidean measure, KM-M shows K-means with Manhattan measure, finally SKM-E(SOM based K-means with Euclidean measure) and SKM-M(SOM based K-means with Manhattan measure) shows the result of the proposed work.

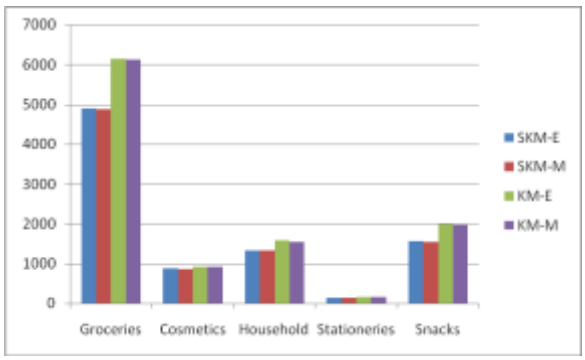

Fig-3: Average Time Parameter

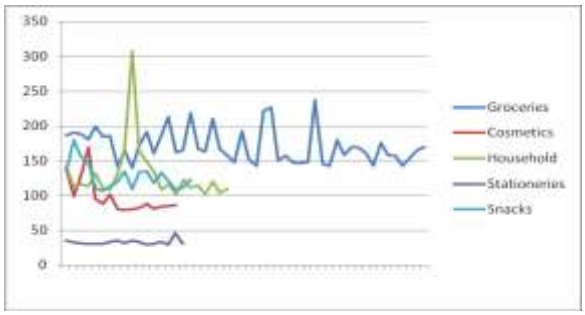

Fig-4: Sales in Super market

The proposed work reveals that the groceries sales is more in the supermarket and the highest product sales is with household item (detergent soap).

VI. Conclusion

This experimental evaluation scheme was created to provide a correct base of performance and also a comparison with other methods. From these experiments on the datasets, it is observed that proposed approach using Self Organizing Map based K-means algorithm has provided correct results in terms of finding frequent purchasing patterns. The proposed approach is helpful in Market Basket Analysis and Stock Management of a super market. At last, the experimental results of SKM are better than the simple K-Means algorithm that has been already used. The average time and error rate are less compared to other methods.

References

[1]. Trybula, W. J. (1997). Data mining and knowledge discovery. In M. E. Williams (Ed.) Annual Review of Information Science and Technology, 32, 196-229. Medford, NJ: Information Today.
[2]. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), 37-54.
[3]. Sharma, S. (1996). Applied Multivariate Techniques. New York: John Wiley & Sons.
[4]. Berry, M. J. and Linoff, G. (1997). Data Mining Techniques. New York: Wiley Computer Publishing.
[5]. Hinton, G. (1992). How neural networks learn from experience. Scientific American, 267(3), 145-151.
[6]. Bacao, F., Lobo, V., & Painho, M. (2008). Applications of Different Self-Organising Map Variants to Geographical Information Science Problems. In P. Agarwal & A. Skupin (Eds.), Self-Organising Maps: Applications in Geographic Information Science (pp. 21-44): John Wiley & Sons, Ltd.
[7]. J.B.MacQueen "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press,281-297 (1967)
[8]. Ackoff, R. F. (1989). From Data to Wisdom. Journal of Applied Systems Analysis, 16, 3-9. Alahakoon, D., Halgamuge, S. K., & Srinivasan, B. (2000). Dynamic Self-Organizing Maps with Controlled Growth for Knowledge Discovery. IEEE Transactions on Neural Networks, 11(3), 601-613.
[9]. N. Sujatha, K. Iyakutty, "Refinement of Web usage Data Clustering from K-means with Genetic Algorithm", European Journal of Scientific Research ISSN 1450-216X Vol.42 No.3 (2010), pp.464-476