Application of Cluster Analysis In Expert System – A Brief Survey

Mamta Tiwari[1], Dr. Bharat Mishra[2]

[1] Dept. of Computer Application, U.I.E.T., C.S.J.M. University Kanpur, U. P., India

[2] Dept. of Physical Sciences, M.G.C.G. Vishwavidyalaya, Chitrakoot. (M.P.) India.

Abstract

This is era of knowledge and information. One very major task that has been evolved now a day is to mine a knowledge base. On the other hand expert systems are used extensively in many domains. There are many applications of expert systems for predicting and finding a feasible solution for any particular problem. Various tools also have been evolves for upgrading and modifying the existing expert systems and making them more useful in their intended purposes. The current paper explains the expert systems that use cluster analysis as a tool and briefly discusses few such expert systems.

Keywords: *Clustering methods, Expert systems, MovieGEN, Illiad, Tourist Expert systems.*

1. Introduction

There is virtually an explosion of information these days but unfortunately this tremendous progress of mankind in every walk of life is largely concentrated in the urban limits. In a country like India and the countries of the so-called third world, the importance of good expert system is beyond description.

Expert systems are intelligent computer programs that are designed to simulate the problem-solving behavior of a human being, who is an expert in a narrow domain or discipline. Since the human race is still facing plethora of problems directly related to their life and livelihood involving poor living conditions, poor medical facilities, lack of educational facilities and recreational facilities etc. Expert systems can be proved a boon in disguise for such sufferers. There are many applications of expert systems ranging from agriculture, finance, education, medicine to military science, process control, space technology and engineering.

Extracting knowledge from existing sources of information is a key development area to unlock yet unknown relationships between specific data point and data mining can be proved as great help in this regard.

Before discussing applications of various tools of data mining especially cluster analysis in expert system, let us have a look on clustering and various methods and techniques used for clustering.

The process of grouping or making sets of nearly similar type of physical or abstract objects is known as clustering. The groups thus formed are known as clusters. Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters [1]. We can compare the clusters with classes as in object-oriented programming paradigm. The slight difference between clusters and class is that, in class every object of it is exactly identical in properties whereas in cluster, every object is similar to other objects of its cluster and dissimilar to the objects of other cluster based on some particular properties.

When data mining is concerned, clustering is having an edge over classification. In data mining, we have to mine a large set of data, clustering saves us from a costly overhead of collection and labelling of a large set of training tuples or patterns, which the classifier uses to model each group. Clustering is sometimes also called as data segmentation in some applications because clustering partitions large data sets into groups according to their similarity.

There are several clustering techniques available and those are organized into the following categories as partitioning methods, hierarchical methods, density-based methods,

grid-based methods, model-based methods, methods for high-dimensional data and constraint-based clustering.

2. Clustering Methodology

We here briefly present various methods of clustering techniques [1].

2.1 Partitioning Method

Given a database of *n* objects or data tuples, a partitioning method constructs *k* partitions of the data, where each partition represents a cluster and $k \leq n$. That is, it classifies the data into *k* groups, which together satisfy the following constraints:

(1) Each group must contain at least one object.
(2) Each object must belong to exactly one group.

Some popular techniques as k-mean and k-medoids are placed in this category.

2.2 Hierarchical Method

A hierarchical method creates a hierarchical decomposition of the given set of data objects. It can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed.

The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into one or until a termination condition holds.

The divisive approach which is also known as the top-down approach starts with all of the objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds. Chameleon and BIRCH are some techniques of this kind.

2.3 Density-based Methods

Most partitioning methods cluster objects, based on the distance between objects. Such methods can find only spherical-shaped clusters and encounter difficulty at discovering clusters of arbitrary shapes. That is why some other clustering methods have been developed those are based on the notion of density. The general idea is to continue growing the given cluster as long as the density (number of objects or data points) in the neighbourhood exceeds some threshold; that is, for each data point within a given cluster, the neighbourhood of a given radius has to contain at least a minimum number of points.

DBSCAN and its extension, OPTICS, are typical density-based methods.

2.4 Grid-based Methods

Grid-based methods quantize the object space into a finite number of cells that form a grid structure. All of the clustering operations are performed on the grid structure (i.e., on the quantized space). The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space. STING is a typical example of a grid-based method.

2.5 Model-based Methods

Model-based methods hypothesize a model for each of the clusters and find the best fit of the data to the given model. A model-based algorithm may locate clusters by constructing a density function that reflects the spatial distribution of the data points.

EM is an algorithm that performs expectation-maximization analysis based on statistical modelling. COBWEB is a conceptual learning algorithm that performs probability analysis and takes concepts as a model for clusters. SOM (or self-organizing feature map) is a neural network-based algorithm that clusters by mapping high dimensional data into a 2-D or 3-D feature map,

2.6 High - Dimensional Data-based Methods

Many applications require the analysis of objects containing a large number of features or dimensions. As the number of dimensions increases, the data become increasingly sparse so that the distance measurement between pairs of points become meaningless and the average density of points anywhere in the data is likely to be low. CLIQUE, pCluster and PROCLUS are some such techniques.

2.7 Constraint-based Method

This is a clustering approach that performs clustering by incorporation of user-specified or application-oriented constraints. A constraint expresses a user's expectation or describes "properties" of the desired clustering results, and provides an effective means for communicating with the clustering process. Various kinds of constraints can be

specified, either by a user or as per application requirements.

3. Various Application of Clustering in Expert Systems

We have chosen for review, three expert system from three different areas of utilization and interest. The one we first pick for the review is Iliad [2]. Iliad was designed to solve a broad variety of internal medical problems. The others are MovieGEN [3] and an expert system for tourism. MovieGEN, an expert system for movie recommendation. The system uses machine learning and cluster analysis based on a hybrid recommendation approach. The system takes in the users' personal information and predicts their movie preferences using well-trained support vector machine (SVM) models [3].

A brief review of the above written three types of expert systems is given next, starting with the Iliad.

3.1 Iliad

Iliad is a medical expert system designed to run on Macintosh computers. Iliad is a expert system whose medical knowledge is organized by disease into "frames" that each contains multiple findings that may be expected in that disease [2].

Most medical expert systems contain two essential components: an "inference engine" and a "knowledge base". Iliad's knowledge base was frame-oriented and was upgraded by including clustered knowledge frames. Clusters are Boolean decision frames that contain finding those are conditionally dependent and describe pathophysiologic concepts.

Frames in Iliad, contain multiple findings that may be likely to present in that disease. These findings are processed sequentially, using Bayes' Theorem, when knowledge about the patient becomes available. Iliad incorporates newly designed knowledge frames called "clusters".

The clustered model of knowledge representation was developed for Iliad because the previous, non-clustered model caused Iliad to produce inaccurate, overconfident diagnoses. Since Iliad's inference engine uses Bayes' theorem, all patient findings used in a case must be, conditionally independent but the assumption of conditional independence required by Bayes' theorem is often violated in medicine. For instance, the patient findings "fever" and "chills" commonly co-occur. Such conditionally dependent findings, if used together in a Bayesian frame, bias the inference engine towards overconfident diagnoses.

A new, "clustered" knowledge model was devised to solve the problem of conditional dependence. Clusters are knowledge frames containing highly conditionally dependent findings. These groups of findings often have pathophysiological meaning. Most clusters are decided according to Boolean logic. Boolean frames may be decided with various levels of certainty (e.g., "definite" "probable" "unlikely" or "absent"). Decisions made about clusters are passed to Bayesian frames, much as procedures in Pascal pass variable results to the main program.

The cluster analysis was performed using the BMDP8M (BioMeDical Proprietary package version 8M) cluster analysis programs on the University of Utah's Sperry Univac 1100 mainframe computer. Although the goal of cluster analysis is to generate groups of highly similar findings, this goal can actually be reached most easily by starting with a matrix of dissimilarities between the findings. Groups of highly similar findings can be defined as having low dissimilarities. The elements of this dissimilarity matrix can be based on Pearson correlations $(1 - r^2)$, conditional probabilities [1 - p(fi/fj)], or Euclidean distances $(fi - fj)^2$.

The final step in the cluster analysis program is assembly of a "results" matrix. This is a two-dimensional matrix of findings by dissimilarity scores. The findings dimension must be rank ordered by average dissimilarity. The BMDP8M program has several strategies to accomplish this goal. One easily understandable way is to locate the most dissimilar finding and set it aside. This process can be performed recursively until all the findings are sorted by dissimilarity. This process produces a results matrix that is rank-ordered along the findings dimension.

There is evidence from psychological research that humans naturally employ hierarchically structured, clustered knowledge models. Cluster analysis is a technique that can be used to discover and validate clustered knowledge concepts. Since clustered knowledge models are natural for humans, a clustered expert system may be able to provide better explanations for its diagnoses.

3.2 MovieGEN

MovieGEN, an expert system for movie recommendation. The system was implemented using machine learning and cluster analysis based on a hybrid recommendation approach. The system takes in the users' personal

information and predicts their movie preferences using well-trained support vector machine (SVM) models [3].

Based on the SVM prediction it selects movies from the dataset, clusters the movies and generates questions to the users. Based on the users' answers, it refines its movie set and it finally recommends movies for the users.

Recommendation systems are special types of expert systems in the sense that they combine the knowledge of the expert in a given domain (for the product type being recommended) with the user's preferences to filter the available information and provide the user with the most relevant information. Personalization of the recommendations works by filtering a candidate set of items (such as products or web pages) through some representation of a personal profile. Two main paradigms for the filtering are content-based approach and collaborative approach. Most recommendation systems use a hybrid approach, which is a combination of these two approaches. A content based recommendation system uses the user's past history to recommend new items where as a collaborative approach uses the preferences of other people with similar tastes for recommending items to the user.

The MovieGEN, a movie recommendation system was developed that recommends movies to users based on their personal information and their answers to questions based on movies. A user model was created using SVM based learning techniques. Using this model it can be predicted the genres and the period of the movies that the user prefers based on the user's personal information.
This incorporates the collaborative approach i.e. the user's choices are predicted based on the choices of other similar users. A variation of the content based approach was implemented by taking into consideration the user choices not based on the user's past history but based on the answers he gives to the questions asked by the system. The system has been developed in Java and currently uses a simple console based interface. Machine learning constitutes an essential step in this approach. For any machine learning model, the data sets are composed of two parts, the input and the output. The output is usually the subjects of interest, in other words, the part that we want to predict or classify, while the input is the set of elements that might have impacts upon the output. Machine learning attempts to correlate the output and the input, by approximating functions in between whose formulations are unknown.

K-means is used in this approach as the cluster analysis tool. K-means is one of the most widely-used partitioning methods in the data mining community, and has been studied and applied in a wide range of domains, including bioinformatics (Guralnik and Karypis, 2001; Zhong et al., 2005), pattern recognition (Estlick et al, 2001; Saegusa and Maruyama, 2007; Filho et. Al, 2003), text classification (Steinbach, 2000), etc.

Support Vector Machine (SVM), an effective and efficient machine learning tool that has been extensively studied within the machine learning community, is utilized in this expert system as the machine learning algorithm. SVM is incorporated in this system to establish a
correlation analysis between personal particulars of a user and his or her personal preference for movies. For each set of user input, a SVM is trained based on a predefined set of training samples, which increases in size after each time the system is used.

In this system, there are two sets of data, the training data set and the testing data set. The training data set comprises of different training samples, each of which is a combination of an input vector and an output vector. The testing data set comprises of different testing samples, each of which contains only an input vector, while the output vector is to be predicted by the machine learning.

Once the genres and period the user prefers based on his/her personal information using SVM, has been predicted, this information can be used to select movies from the dataset, generate questions about these movies and finally return a refined movie recommendation to the user.

3.3 An Expert System for Summer Tourism

This study had an aim to support tourism sector in Turkey by using an expert system [4]. Thus, tourists will be able to select the most suitable holiday places for themselves. Before the tourists go to a holiday place which they have not visited before, they make a research about this place. Also, some surprises in this place are learnt before the tourists go and many tourists do not like this situation.

Therefore, an operation of text mining was preferred in this study. Thus, tourists do not need a research about the holiday places. All that the expert system will return will be a decision according to users' preferences. The expert system had an aim to return more decisions than one. When a tourist uses the system; only one place is not returned, sorted places from the most suitable place to the least suitable one are given. Therefore, a clustering structure was needed. After the system decides the most suitable place for the tourist; the cluster where this suitable place locates finds and the all holiday places in this cluster

are recommended in order from the most suitable to the least suitable.

There are lots of features as attributes from text collection although there will be a low number of holiday places; thus, a large dataset was obtained. Therefore, K-Means clustering algorithm as both simple and fast clustering algorithm was preferred. However, K-Means has problem about deciding the space of clusters, because K-Means can give a different space of clusters with same dataset at each working. The cause of this situation is that K-Means starts clustering with random initial centre points. Therefore, K-Means++ clustering was used as a new approach to K-Means without random initial centre points and with consistent result spaces.

This study had four steps briefly. Firstly, the most preferable places for summer holiday in Turkey were decided. According to a research on web pages of Cultural and Tourism Ministry of Turkey about tourism, the most important places are Alanya, Ayvalık, Bodrum, Çeşme, Datça, Didim, Dikili, Fethiye, Kaş, Kuşdası, Marmaris, Side and Yalova. These places are preferred by both foreign and regional tourists a lot because of both common and unique features of these places. Therefore, secondly, the features must be determined. For this step, a research with rich documents about these places was done on web and these documents were collected in a text file for each place. These text files would be used for text mining operations in the next steps.

Thirdly, a dictionary was created for each place from the collection of text files. These dictionaries are too large to process, because these dictionaries content stop-words and unnecessary words for tourism. Therefore, some words were determined to be deleted from the dictionaries and they were deleted; thus, the satisfactory dictionaries were obtained for each holiday places. A data warehouse must be needed for mining operations on these dictionaries.

Therefore, pre-processes with vector space model were needed; thus, a dataset was obtained with tuples and their attributes. In last step, this dataset was used by K-Means++. It gave a space of clusters where there were the places. Finally, an expert system was ready to use and holiday places were recommended according to these clusters and the expectations of tourists.

4. Summary and Future Scope

We strongly believe that clusters and cluster analysis should be a part of expert systems because they can improve the accuracy of Bayesian decision systems [1]. But clusters are also important because they model innate human knowledge structures. Strong evidence supports our assertion that cluster-like knowledge structures are natural human mechanisms for organizing information. While mechanisms of human knowledge organization are incompletely understood, it is clear that humans must employ simplifying heuristics in complex situations. The cluster heuristic allows data to be combined into useful patterns that may lead to better decisions.

In present scenario the application of expert system has already gained momentum, still there are lot of areas where a great deal of efforts is still required. The knowledge engineers and information scientists have done tremendous work if form of expert systems now there is an immense need to upgrade them and make them more useful. We believe that various data mining approaches and techniques such as k-mean, pCluster and STING etc. are going to play a vital role in this mega job. The validation and improvement of expert system, which in turn is already a very complex phenomenon, is the demand of time and we hope that we can see equal efforts in this direction also.

5. References

[1] Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques 2nd Ed. - Morgan Kaufmann Publishers

[2] Michael J Lincoln, Charles Turer, Brad Hesse MS, Randolph Miller - A Comparison of Clustered Knowledge Structures in Iliad and in Quick Medical Reference.

[3] Tilak Gaurangi, Eyrun A. Eyjolfsdottir, Nan Li MovieGEN: A Movie Recommendation System.

[4] Yunus Doğan and Alp Kut - An Expert System for Summer Tourism in Turkey by Using Text Mining and K-Means++ Clustering ,ICT Innovations 2010 Web Proceedings ISSN 1857-7288

Mamta Tiwari completed her Master of Technology in Computer Science from U.P. Technical University, Lucknow (U.P.) in 2006. Earlier she had completed Master of Computer Applications from Rani Durgavati University, Jabalpur (M.P.) in 2001 and Master of Science from Kanpur University, Kanpur (U.P.) in 1994. She has put up over 10 years of teaching experience to engineering students. Presently she is working as a lecturer in the Dept. of Computer Application in University Institute of Engineering and Technology, C.S.J.M. University, Kanpur. Presently she is perusing her Doctoral Program from Mahatma Gandhi Gramodaya Viswavidyalaya, Satna (M.P.) India. Her research interest includes Data Mining, Artificial Intelligence and Software Engineering.