Comparing Models for Analysing Database Pattern

Lakshya Goyal

Wynberg Allen School
Mussoorie, India
emlakshya@gmail.com

Abstract—In today's scenario, size of database is growing at a tremendous speed and analyzing such data for various purposes is of utmost importance. In this paper, we have applied our methodology to datasets of different sizes and discussed the experiment results in analyzing the pros and cons of various models. We have given an implementation model for knowledge discovery from huge unlabeled temporal databases by employing a combination of HMM and K-means technique.

Keywords - pattern analysis, HMM, KMM, KM HMM, recursive model

I.INTRODUCTION

With the huge growth of data [1], there has been a pressing issue to find tools and techniques to extract meaning information from this raw data and convert it into knowledge. For this, the data must be analysed by finding the patterns of the raw data and thereafter use these patterns to generate knowledgeable data.

These can be achieved through data analyses, which involve simple queries, simple string matching, or mechanisms for displaying data. The main aim of studying the various techniques of classification is the development of a tool or algorithm which can be utilized for prediction of the class of an unknown object, which is not labelled. This tool or algorithm is called a classifier while the objects in the classification process are represented by instances or patterns, where a pattern comprises of several attributes. The classification accuracy of any classifier is assessed by how many testing patterns it is able to classify correctly [2].

In general, unlabelled data is mined with the help of data clustering. Clustering, as a robust tool of Knowledge Discovery, aims to find hidden patterns in datasets by grouping data items together according to some criterion of closeness [3,5,13]. Grouping of objects is required for various purposes in various fields such as science and technology, social sciences, medical sciences, etc.

Studies have discussed various performance metrics and parameters are provided, and a comparative assessment of the corresponding aspects like cluster head selection, routing protocols, reliability, security, and unequal clustering [4].

The functionality and effectiveness of KM HMM [9,10] recursive model is first tested on one subset and its results are then backed up by different subsets of larger sizes. KM HMM [9] stands for the model-based clustering approach. It combines the power of K-means clustering with HMM, where K-means initializes the experiment by clustering the profiles. The resulting clusters are then used as an auto-labelling mechanism such that profiles that are grouped into the same cluster receive the same label.

The KM HMM model can group unlabelled data according to its underlying structure, i.e., patient's medical behaviours. The discovery is a data driven process which progresses through different hierarchical paths in unearthing the patterns in different age cohorts. Those patterns carry patient's medical information, and a proper interpretation could reveal some hidden but important knowledge such as a medical behaviour pattern of aged diabetics or abnormal medical behaviour pattern which could lead in fraud investigation. These are only few examples of application that can be achieved if some ground truth is made available.

We implement K-means clustering [11] and Hidden Markov model [6] approach, referred as KM HMM[13] recursive model, on smaller subsets in order to obtain a first insight into the effectiveness of the approach [13], and to serve as a basis for comparisons when subjecting the same method to the entire set of data. Then, the findings are then combined to produce the best performing approach which is then applied to the entire set of data.

II. LITERATURE REVIEW

Several clustering algorithms are proposed by various researchers, like Partitioning clustering algorithms, such as K-means [8] and CLARA [5] that assign objects into k clusters(predefined cluster number), that further iteratively reallocate objects to improve the quality of results obtained from clustering. Although K-means is a popular clustering algorithm [6] which is easy-to-understand also but at the same time it is very sensitive to the selection of the initial centroids and has no general solution to find the optimal number of clusters for a given data set.

Few authors have proposed Model-based clustering methods that are based on the assumption, that data is generated by mixture of underlying probability distributions

and optimization with the help of models such as statistical approach, neural network approach and other AI approaches. The typical techniques in this category are AUTOCLAS [19], DENCLUE [7] and COBWEB [20] but they are facing a challenge of choosing a suitable one from the model-based candidates. Clustering, CH selection, routing, reliability and security aspects are addressed in review studies on clustering approaches [2,4]

These clustering-based approaches suffer from high computational cost, especially when the scale of data is very large. Another very commonly used model is the HMM model, the Hidden Markov Model (HMM) which is based on statistical modelling. Encoding of temporal pattern has made this approach very popular. Several variants of HMMs [7,8,9] exist like discrete HMM, continuous observation HMM, and input-output HMM, to name a few but they are also facing the challenge of heavy computational cost. To overcome these issues, we have proposed a recursive model which reduce the computational burden and to improve the quality of the model.

III. PROPOSED MODEL

In this section, our KM HMM approach is detailed through the application on one subset, then its functionality is further confirmed by applying the method on few other subsets. These preliminary experiments are important and necessary since they provide a vehicle to unfold the methodology. The step-by-step introduction of our method, the detailed discussions at the end of each step and the walk through of the sample profiles form the key in explaining the methodology.

3.1 Subset Selection

There are numerous ways in selecting reasonable subsets. What matters here is data quality: the selected subset has to be an unbiased representative of the data. We have divided the whole set of profiles into nine age cohorts, we decided to select one of the age cohorts to carry out our first experiment. Patients in age cohort 45-55 are well away from the female reproductive ages, and are still reasonably youthful as not to suffer from age related illnesses. Their profiles are not expected to be too complex to overly challenge the methodology. The data set is further downsized by constructing a rule-based selection e.g. particular illness based or gender.

IV. RESULTS

The experiment conducted for the study and the results of the experiment for comparing the accuracy of the models are presented in the following sections.

3.2 KM HMM Based Recursive Model

Using a fixed number of clusters for age cohort 45-55.

KM HMM [9] combines the power of K-means clustering with HMM, such that profiles that are grouped into the same cluster receive the same label, whereas profiles in different cluster receive a different one. Thus, the profiles are labeled according to the cluster membership. While these labels do not carry any meaning (other than cluster membership), it allows the application of a supervised learning scheme such as HMM.

By creating a model (HMM) for each cluster, HMM learns to detect patterns in given time series data which best describe the data in a given cluster. The result is a set of HMMs, one for each cluster. The procedure can then be applied recursively to each of the pattern classes in order to further segment a dataset into ever smaller classes. In practice, our methodology is unfolded through the following two steps:

Step 1: K-means Clustering

The first step of the proposed pattern discovery methodology addresses the clustering of data. In our case, the data subjected to the experiments are patients' profiles. K-means [11] clustering takes each profile as a 352-dimensional vector and calculates its distance towards centroids of each cluster, and eventually assigns the profile to the cluster where the distance between the profile and the centroid reaches minimum.to which a person believes that using a particular system would enhance performance.

K-means takes the initial parameters and performs data clustering. The cluster detected by K-means are then analysed, and clusters smaller than 200 are discarded (by assigning the patterns to the remaining clusters). The application of K-means algorithm results in 6 clusters. The patterns in each cluster are then uniquely labelled as Cluster1, Cluster2, ..., Cluster6 respectively. The benefit here stands for the total benefit paid for a patient throughout the year where the profile is drawn against

Table I. Basic information for k means clustering algorithm

Name of Cluster	Number of Profiles	Benefit Range in $
Cluster 1	28,608	8.15-1181.85
Cluster 2	9,427	124.95-2303.65
Cluster 3	1,447	357.25-7894.00
Cluster 4	1,199	393.60-7159.40
Cluster 5	1,216	505.10-9070.65
Cluster 6	903	645.60-8232.35

Step2: HMM Data Modelling and Recursive Mining

With HMM modelling [6], data is assessed according to contextual information embedded within the temporal sequence which we refer to as profiles. The process of esti-

mating the values of these parameters is considered as the training of HMM models. For this experiment, a maximum 1,500 profiles are selected when the cluster size is over 2,000 (such as Cluster1 and Cluster2), otherwise 90% of the data from the cluster are subject to training (such as Cluster3 to Cluster6). Table II specifies the size of each training set.

Table II. Information of dataset in terms of its size in training hidden markov model

Name of Cluster	Number of Profiles	Size of Training Sample
Cluster 1	28,608	1,500
Cluster 2	9,427	1,500
Cluster 3	1,447	1,302
Cluster 4	1,199	1,079
Cluster 5	1,216	1,094
Cluster 6	903	812

For our experiments, we used the freely available, and well matured HMM software package known as HTK version 2.0. We customized the software package to allow the dealing with very large and zip compressed databases. This was necessary since HMM is not normally suited to deal with data mining tasks. One of the parameters that needs to be set when using the HTK 2.0 software package is the number of states in the HMM.

Following experiments are based on a left-right HMM with 3 states. Table III provides the training results of the 6 HMMs in terms of mean and variance of each HMMs. However, experiments on HMMs trained with more than one state were aborted due to lack of information.

Class1 is the largest which attracts 27,740 profiles while Class5 is the smallest with only 186 profiles. It is observed that the K-means clustering algorithm and the hidden Markov model grouped the profiles differently. Upon closer inspection of the K-means clustering results when compared to the HMMs classification results, it can be stated that the HMM classification makes more sense.

On a global scale the K-means clustering algorithm grouped the profiles as a whole profile, it works on the distance from the centroids to point of 352-dimensions. Hence at times it has grouped profiles which to the naked eyes appear to be quite different.

On the other hand, the hidden Markov model groups profiles together based on the time evolution of the profiles. The classification of profiles is the result of considering the context in which benefit is paid. For example, the hidden Markov model can assess whether a benefit paid is out-of-the-ordinary by considering the context within which the benefit was paid, this is an interesting observation given that the HMMs were trained on data labelled by K-means.

This process should keep on continuing recursively till further classification is possible. The process of Recursive mining allows the classes to be re-clustered into sub-clusters where the iterative mining will operate on each of the sub-clusters. The sub-clusters will then be modelled and iteratively mined by HMM so that new HMMs or patterns are discovered for each of the newly generated sub-clusters.

Table III. HMM training results: mean and variance of HMM

Name of Cluster	Number of Profiles	Size of Training Sample
HMM1	6.128910e+00	5.203418e+02
HMM2	2.477937e+01	3.275482e+03
HMM3	4.600096e+01	1.269653e+04
HMM4	5.877817e+01	2.111918e+04
HMM5	6.670536e+01	2.268726e+04
HMM6	7.352632e+01	2.508250e+04

Table IV. Classification result by the hmm model with comparison to k-means clustering

Cluster	CLASS 1	CLASS 2	CLASS 3	CLASS 4	CLASS 5	CLASS 6	TOTAL
HMM1	25.458	3,055	95	0	0	0	28,608
HMM2	2,274	6,005	1,003	82	10	53	9,427
HMM3	6	642	524	82	33	160	1,447
HMM4	1	246	501	123	56	272	1,199
HMM5	1	256	498	112	46	303	1,216
HMM6	0	92	377	95	41	298	903

Figure 1 shows the recursive process of our methodology. It demonstrates what has been done by the two processes of clustering and recursive mining: all the profiles in the data pool of a particular age group are sub divided and processed to get a further level with the help of clustering, modelling, and iterative mining as was described above.

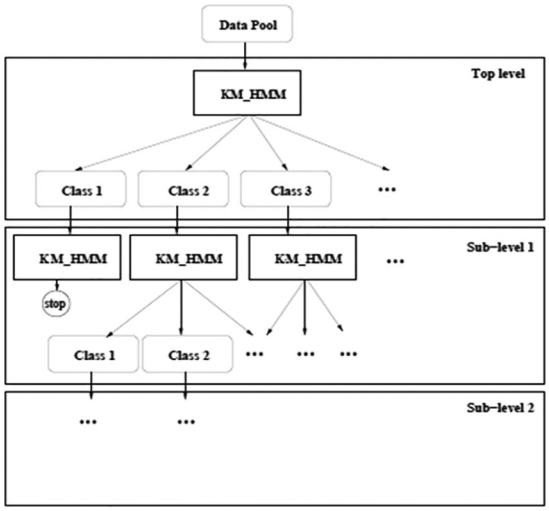

Fig. 1. Recursive refinement

All profiles of the above table of age cohort of 45-55 year are classified by the set of 63 HMMs generated for this cohort group. The classification is performed as fol-

lows: a profile is presented to all 63 HMMs, and the HMM which responds with the highest likelihood value wins. i.e., the profile is classified into the class represents by the winning HMM. Finally, at the end of this recursive modelling, we would like to re-visit the 36 profile which have been discussed in step 1 and step 2.

The 63 final HMMs are applied in classifying these 36 profiles in order to illustrate how these profiles are grouped at the end of the iterative training procedure. It is observed that the final HMMs provide a much finer classification to the profiles.

V. CONCLUSION

In this paper, we have implemented our analytical algorithm i.e. recursive KMM Model on a smaller subset and later on the complete set, in order to understand the effectiveness of KM HMM model. The distinctiveness of our model is the way the clustering KM HMM is applied on the dataset. Initially, to overcome the weakness of HMM on large datasets, we have applied the model iteratively. Also, in place of applying the model on entire dataset in one go by understanding all profiles of training set, we have done the mining iteratively in a controlled manner.

This helps to keep the training time in acceptable limits while keeping the model refined in each iteration. We analysed through the implementation model of KM HMM recursive model that we are getting more refined data as compared to K-means model. Thus iterative process makes HMM available for modelling a large set of data and maintains the known asset and the strength of HMMs. In our further study, we will take larger dataset and find the efficiency of this model and also compare it with other existing models.

The findings of this study are significant for the healthcare service providers as well as the digital marketing companies which may provide the technology and expertise for the same. By considering the various aspects of the factors that affect the attitude and actual adoption, detailed plans can be devised for effective implementation of these technologies. Further research may be conducted in different sectors for establishing the generalizability of these results and qualitative research can be conducted to explore further dimensions of the perceived usefulness and challenges constructs.

REFERENCES

[1] R. Chaiken, B. Jenkins, P. Larson, B. Ramsey, D. Shakib, S. Weaver, and Zhou, 2008, SCOPE: Easy and Efficient Parallel Processing of Massive Data Sets. PVLDB, 1(2):1265–1276. https://doi.org/10.14778/1454159.1454166

[2] Saxena, Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Joo, M., Ding, W., & Lin, C. T. (2017). A review of clustering techniques and developments. Neurocomputing, 267, 664-681.

[3] Zhang, Y. Yin, Y., Guo, D., Yu, X. & Xiao, L. (2014). Cross-validation based weights and structure determination of Chebyshev-polynomial neural networks for pattern classification. Pattern Recognition, 47(10), 3414-3428.

[4] Amutha, J., Sharma, S., & Sharma, S. K. (2021). Strategies based on various aspects of clustering in wireless sensor networks using classical, optimization and machine learning techniques: Review, taxonomy, research findings, challenges and future directions. Computer Science Review, 40, 100376.

[5] L. Kaufman and P. J. Rousseeuw, 1990, Finding Groups in Data: an Introduction to Cluster Analysis, John While & Sons.

[6] S. Kobayashi, T. and Haruyama.1997,Partly-hidden markov model and its application to gesture recognition. In 1997 IEEE International Conference on Acoustics, Speech and Signal Processing, Vol.4, pages 3081–3084.

[7] A. Hinneburg and D. Keim,1998, "An Efficient Approach to Clustering in Large Multimedia Databases with Noise", Proceedings of KDD-98 .

[8] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, 1998, "Automatic subspace clustering of high dimensional data for data mining applications", Proceedings of the ACM SIGMOD Conference, Seattle, WA., pp.94-105 .

[9] G. Sheikholeslami, S. Chatterjee, A. Zhang, 1998, "Wave cluster: A multi-resolution clustering approach for very large spatial databases", Proceedings of Very Large Databases Conference (VLDB98), pp.428-439.

[10] A. Panuccio, M. Bicego, and V. Murino, 2002,.A hidden markov model-based approach to sequential data clustering. In Proceedings of Joint IAPR International Workshops SSPR 2002 and SPR 2002, pages 734–743.

[11] M. P. Perrone and S. D. Connell, 2000, K-means clustering for hidden markov models. In Proceedings of the 7th International Workshop on Frontiers in Handwriting Recognition, pages 229–238.

[12] P. Smyth. 1997. "Clustering sequences with Hidden Markov Models" Advances in Neural Information Processing Systems, Vol.9:648–654.

[13] Babita, Paramjeet Rawat, Parveen Kumar, 2014, "An Approach to Analyze Pattern from Large Database of Healthcare" International Journal of Engineering and Innovative Technology (IJEIT) Volume 4, Issue 1

[14] J. B. MacQueen,1967, "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, pp.281-297.

[15] L. Kaufman and P. J. Rousseeuw,1990 "Finding Groups in Data: an Introduction to Cluster Analysis", John Wiley & Sons.

[16] D. Jiang, A. K. H. Tung, and G. Chen., 2010, Map-join-reduce: Towards Scalable and Efficient Data Analysis on Large Clusters. TKDE, 23(9):1299–1311.

[17] S. Han, D. Chen, M. Xiong, and A. K Mok. 2008. Online Scheduling Switch for Maintaining Data Freshness in Flexible Real-Time Systems. In Proc. of RTSS, pp. 115–124. . https://doi.org/10.1109/RTSS.2009.36

[18] Y. Xiong and D. Y. Yeung., 2002, Mixtures of arma models for model-based time series clustering. In Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02), pages 717–720. https://doi.org/10.1109/ICDM.2002.1184037.

[19] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman, 1998, "AutoClass: A bayesian classification system", Proceedings of 5th International Conference on Machine Learning, Morgan Kaufmann, pp. 54-64

[20] D. Fisher, 1987, "Improving Inference through Conceptual Clustering", Proceedings of 1987 AAAI Conferences, Seattle Washington, pp.461-465.