ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 1, July 2014

An Approach to Analyze Pattern from Large Database of Healthcare

BABITA, PARAMJEET RAWAT, PARVEEN KUMAR
Mewar University, UP Technical University, Amity University

Abstract- To address the problem of knowledge discovery through pattern matching various models are used. There are effective methods for discovering knowledge from temporal data. Like, hidden Markov models (HMM) are a popular approach to discover patterns from temporal data. However, HMMs do not scale favorably with the size of a given dataset, and hence, are not normally used for data mining applications of large sets of temporal data. In this paper, we present a new method in discovering patterns from a large set of unlabeled temporal data. K-mean (KM) and hidden Markov model (HMM) form the core of our proposed approach. These methods are engaged to cluster temporal data by using a novel recursive KM- HMM model.

***Index Terms* – knowledge discovery, pattern discovery, clustering model, KM-HMM.**

I. INTRODUCTION

The work of industrial engineers is to devise the best means of optimizing processes in order to create more value from the system; data-mining [2] has become a powerful tool for evaluating and making the best decision based on records so as to create additional value and to prevent loss. The potential of data-mining in the field of medical has yet to be fully exploited as it plays a vital role in diagnosing the disease and giving right medication.

The discovery or extraction of information in a data store is becoming an increasingly challenging task as the size of data is increasing day by day. Much more challenging is the task of finding patterns [1] in large sets of data. This exercise is commonly referred to as Data Mining [3]. Data mining assumes no prior knowledge about the data in a dataset. In this paper, we have developed anew method in discovering patterns from a large set of unlabeled temporal data. We apply this methodology in finding patterns in a large scale database of logged information from Health Insurance Commission. We have used a combination of K-means and HMM for refining the search and extracting the result.

II. BACKGROUND AND RELATED WORK

K-means [14] is one of the simplest unsupervised learning algorithms. It organizes a set of unlabeled objects (data) into groups (clusters) where the objects within a cluster [4] share certain similarities. The grouping is done through an iterative fashion with the goal of minimizing the sum of squares of distances between data and corresponding cluster centroids [19]. K-means algorithm pre-determines the number of K clusters that can be computed on a given set of data by minimizing the overall distances between data points and centroids.K-means algorithm has been very popular because of its simplicity and the capability of converging quickly.

The Hidden Markov Model (HMM) [13] on the other hand has been widely used in temporal data classification [8] [20], temporal pattern recognition such as in speech recognition [7], handwriting recognition [6], gesture recognition [5]. The proven capabilities of the HMM to encode temporal patterns have made this approach a very popular approach to pattern discovery applications. A number of variants of HMMs exist [20]. These include discrete HMM, continuous observation HMM, and input-output HMM, to name a few. Distance based method implies there is not a strong structural relationship within the data, i.e., the relationship is weak. Algorithms such as K-means, Hierarchical Clustering, Cluster-*C*[17] and SEQOPTICS(SEQuence clustering with OPTICS)[12] belong to this category.

III. PROPOSED MODEL

Our approach is based on model based clustering; it normally requires a clustering algorithm to initiate the training of a selected model. We have chosen-means as it is the most commonly used clustering algorithms because of its efficiency, easiness to use and implement. We have refined the clustering result is refined through recursive model using Hidden Markov model (HMM).The problem with the clustering algorithms is the choice of number of clusters as the algorithm do not have methods to access these numbers and assume the user has *a priori* knowledge. In such circumstances, depending on the task, the number can either be inferred from the dataset itself or has to be suggested based on the user's experience. We first label the data and then try to discover the patterns. The following presents the details of our methodology step by step.

Step 1: Data Segmentation

First of all we segment the large dataset into smaller ones and also group the data according to some of its nature which will assist in the data labeling. Segmentation can be taken from a different angle and at different levels depending on the nature of data and the potential research requirements. We are using health insurance data, so segmentation is done by grouping patients into smaller subsets through a natural divider such as age or gender. From the research perspective of pattern discovery, age would shed more light than gender

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 1, July 2014

since most of treatments are not gender specific. Age segment the data forming more subsets which is very much needed in the effort of breaking up the large dataset. Patients of similar ages would have similar medical behaviors, such as young boys of age 7 to 15 are generally healthy and sporty, and likely suffering sport injuries, while girls of age 16 to 24 (or alike)are in their reproduction ages. Therefore, instead of segmenting the patients into over 100 groups, patients are grouped into nine age cohort. We divided our data into nine different age groups i.e. 0-3, 4-6, 7-15, 16-24, 25-35, 36-44, 45-55, 56-70, and 71-110.

Step 2: Data Clustering using K-means or Gaussian Mixtures

Given the nature of our dataset, we believe there are more than one patterns/models in each of the nine datasets. i.e., not all the patients in same age cohort, say 25 to 35, have the same medical pattern, some are sicker than others.

Unfortunately, there is not much *a priori* knowledge about the data within each age cohort. Nevertheless, it is reasonable to believe that there are more than one medical patterns, i.e., more than one labels existing in further description of data nature. Through K-means (KM) cluster can be considered as a process of labeled such that data in the same cluster shares the same label. In general, data clustering does not require much information about the meaning of the data, the only parameter that user has to provide is the number of clusters, i.e., the number of sub-groups the data belongs to.

Step 3: Data Modeling using Hidden Markov Model (HMM)

These labeled cluster can be used for data modeling such that one model is expected to be discovered for each clusters.HMMs has been proved to be successful in mining temporal data where data is context dependent or context sensitive. The patient's profiles which will be used in the training of HMMs are medical history records therefore they appear in certain order to make sense in terms of medical procedure. i.e., certain context dependency exists with profiles. For example, a "subsequent visit to an orthopedic" has to follow "initial visit of an orthopedic", not the other way round. The outstanding advantage of HMM is that it takes the context of the temporal profile into account. Thus makes it an ideal tool for this pattern discovery task. The challenge in the application of HMM is the training cost: HMM could be very expensive if the size of the training set is large. But our methodology is designed in such a way that it will mine the dataset (profiles) for patterns in an efficient and effective manner by employing a training set with decent but controlled size. A threshold or a minimum number of profiles is needed to guarantee the quality of HMM.

The result of data modeling in each pool and for each age cohort respectively, is a set of HMMs which are trained in a controlled manner based on the set of clusters available for that age cohort. These HMMs are considered as the medical behavior patterns discovered for the age cohort and assumed as representative to all the profiles in the age cohort.

Step 4: Iterative Mining

This section is to discuss how to overcome the limitation caused by the size control of HMM training sets to improve the mining quality. We propose this iterative training of HMMs as the solution to strengthen the HMMs.

First of all, we introduce the notation used through our *Iterative Mining*.

• [i]R<j>-HMMs stands for the set of HMMs obtained at the end of the *j*th (j > 0) iteration, irefers to the level which will be discussed in next section. *i*, is optional when *i*=0.

• [i]R<j>-Classes stands for the set of classes generated at the end of the *j*th (j > 0)iteration, *i* refers to the level which will be discussed in next section. *i*, is optional when *i*=0. When j = 0, [i] R0-Classesare in fact [i] R0-Clusterse.g., R36-HMMs are the set of HMMs obtained at the end of the 36th iteration, and R36-Classesare the classification results of R36-HMMs.

The iterative mining unfolds this way: for a given data pool of a certain age cohort, profiles are clustered first, and the clusters generated are referred as R0-Clusters. And then the profiles will be modeled. The number of profiles used to train a HMM is capped. The cap value (the size of the training set) should not be too small to allow a decent training of the HMM, and should not be too large to avoid the curse of the computational complexity of HMM. On the other hand, a bottom line is also needed to draw up to ensure the quality or robustness of HMM. If the total number of valid clusters is less than 2, recursion terminates. For clusters which are large enough, a HMM will be trained for each of them. And the HMMs will be referred as R1-HMMs (the first iterative execution). Trained HMMs can be considered as the discovered patterns, i.e., they are the models which represent the medical behaviors of the patients in the clusters. And when these R1-HMMs are used as classifier, the profiles in the data pool of the age cohort (including the profiles from the cluster where no HMM was trained due to lack of profiles in that cluster) will be classified into classes, which could be referred asR1-Classes. The algorithm stops if convergence has occurred, the size of the clusters fall below a given bottom line, or until a set number of iterations have been reached. More specifically, *convergence* is defined as follows:

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 1, July 2014

• Training the models is said to have converged if one of the two situations are reached:
(1) HMMs in the current iteration generate the same classes as HMMs in the immediate previous iteration. This indicates that HMMs have reached a local convergence.
(2) HMMs yields only one valid class. i.e., the profiles can be modeled by one single HMM. In this case, any further iterative training would not change the fact the profiles share the same pattern.

The set of HMMs are considered as the best model for the data, and the classes' are the best classification so far. The following notation is used to refer the results.

• *[i] FS-HMMs* stand for the final set of HMMs achieved at the end of the iterative mining. i, refers to the level which will be discussed in next section. i, is optional when *i*=0.
• *[i]FS-Classes* stands for the final set of classes generated at the end of the iterative mining. *i* refers to the level which will be discussed in next section. *i* is optional when *i*=0.

The iterative mining approach can be illustrated as shown in following algorithm. The first round training of HMMs is then guided by these R0-Clusters producing a set of trained HMMs: the R1-HMMs. These are then used to conduct data classification yielding a set of classes: the R1-Classes.

ALGORITHM

1. Perform Clustering.
2. If valid cluster>1, then
    a. Perform Modeling : HMM training
    b. Perform Labeling or Classification of Data
    c. If training the model is converged, then
       Goto Step 3
       Else
       GotoStep2 (a)
3. STOP

Step 5: Recursive Refinement

Recursive mining opens this opportunity by allowing 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 modeled and iteratively mined by HMM so that new HMMs or patterns are discovered for each of the newly generated sub-clusters as shown in the following figure.

The recursive refinement at Sub level 1 is completed when all the classes in the Top level are processed. At the end of this sub-level, a new set of HMMs, which is a collection of all the HMMs trained based on each class from the previous level (i.e. the Top level), is obtained along with a new set of corresponding classes. The recursive mining continues to refine the models in the next sub-level until no further refinement is possible. i.e., either it converges or not enough profiles to proceed.

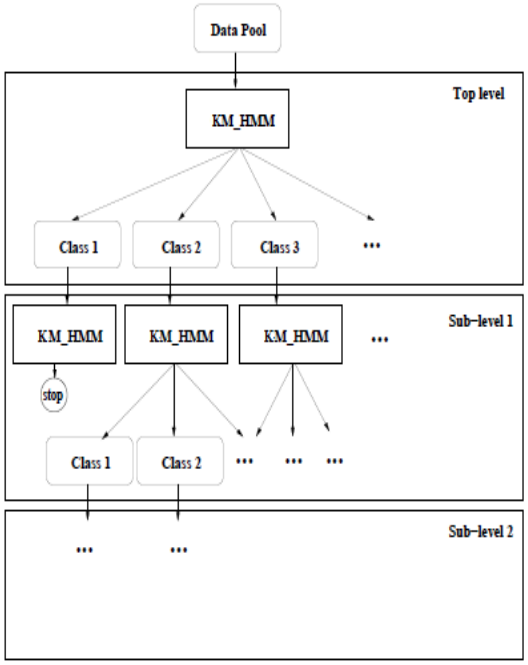

Fig 1.Recursive Refinement

Step 6. Patterns Discovered

The final set of HMMs/patterns is the collection of all HMMs (at the leaf level) where the corresponding classes do not need further recursive mining.

IV. CONCLUSION

A novel recursive hierarchical approach has been introduced to tackle the two research issues of our temporal data mining task: the large size and the lack of labeling. The core of our methodology is a model based clustering which is enabled by a data clustering and a data modeling processes. Model based clustering has been applied in mining temporal datasets thus is not new. The novelty of our method is how the model based clustering KM HMM is applied. First of all, the modeling has been implemented iteratively to overcome the weakness of employing HMMs on large datasets. Instead of modeling the data in one go by learning all the profiles in the training set at once, iterative mining is employed to model patterns in a controlled manner so that the training times of HMMs at each iteration is kept within acceptable limits while the models are more and more refined through the iterations. Thus iterative process not only makes HMM available for modeling a large set of data but also maintains the known strength and the robustness of HMMs.The final set of HMMs are a model of the patterns discovered.

REFERENCES

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

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 1, July 2014

[2] Abdelmelek, S.B. Saidane, S. Trabelsi, M. Base oils Biodegradability Prediction with Data Mining Techniques, Algorithms 3:92-99, 2010.

[3] J. B. Mac Queen, "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 (1967).

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

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

[6] V. S. Nalwa. Automatic on-line signature verification. Proceedings of the IEEE, Vol.85 (2):215–239, 1997.

[7] L. Rabiner and B. H. Juang.Fundamentals of speech recognition. Prentice Hall, 1998.

[8] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman, "Auto Class: A Bayesian classification system", Proceedings of 5th International Conference on Machine Learning, Morgan Kaufmann, pp. 54-64 (1988).

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

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

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

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

[13] A. Panuccio, M. Bicego, and V. Murino.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, 2002.

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

[15] P. Smyth. Clustering sequences with hidden markov models. Advances in Neural Information Processing Systems, Vol.9:648–654, 1997.

[16] J. G. Deller, J. R. Jr.andProakis and J. H. L. Hansen. Discrete-time Processing of Speech Signals. MacMillan Publishing Company, 1993.

[17] Z. X. Ying and J. H. Chiang.Pattern discovery on complex diagnosis and biological data using fuzzy latent variables. In IEEE 23rd International Conference on Data Engineering, 2007. ICDE 2007, pages 576–585, 2007.

[18] M. Athanassoulis, S. Chen, A. Ailamaki, P. B. Gibbons and R. Stoica.MaSM: Efficient Online Updates in Data Warehouses. In Proc. of SIGMOD, 2011.

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

[20] A.R. Post and J.H. Harrison.Temporal data mining. Clinics in Laboratory Medicine, 28(1):83–100, 2008. S. Han, D. Chen, M. Xiong, and A. K Mok. Online Scheduling Switch for Maintaining Data Freshness in Flexible Real-Time Systems. In Proc. of RTSS, pp. 115–124, 2009.