Alternate Medicine Recommendation using Data Mining

Ankita Rewade, Dr. S.W.Mohod

Department of Computer Engineering BDCE Wardha

Abstract - On the Internet, where the number of choices is overwhelming, there is need to filter, prioritize and efficiently deliver relevant information in order to alleviate the problem of information overload, which has created a potential problem to many Internet users. Alternate Medicine System solve this problem by searching through large volume of medicine information to provide users with filter and services. This project explores the different characteristics and potentials of different recommendation techniques in recommendation systems in order to serve as a compass for research and practice in the field of medical recommendation systems. In this project we have used a dataset over 100+ medicines from different companies and brands to do recommendation based on content of medicine and then filter it based on rating and cost based analysis.

Keywords: Random Forest, Data Mining

I. INTRODUCTION

Alternate Medicine System research has made significant advances over the past decades and has seen wide adoption in electronic commerce. Recently, a variety of types of side information (e.g., social friends, item content) has been incorporated into Alternate Medicine System to further enhance their performance, especially the well-recognized problem of data sparsity. However, most of existing approaches have only investigated the value of a single type of side information at a time, such as social trust, friendship, or item contents.

It is necessary to build new theories, techniques and methods to exploit multi-dimensional (homogeneous and heterogeneous) side information to provide users with better personalized recommendations. At the same time, the large volume and variety of side data and the velocity of incremental updates in live systems provide challenges for the scalable mining and application of user preferences.

It is evident that the health of an individual significantly affects her quality of life. For this reason, finding appropriate physicians to diagnose and treat medical conditions is one of the most important decisions that a patient must make. Currently, patients have two options that can aid them in addressing this problem, but both are of limited applicability. The first option is to rely on friends and family for advice on where to seek treatment. While recommendations produced by a close circle of friends can be assumed to be very trustworthy, the likelihood that friends and family have experience with the same medical history as the patient is quite low.

Furthermore, such advice can often be unavailable when, for instance, a patient moves to a new area and does not have an established network from which to seek advice; even when this is not the case, the number of physicians which friends and family have had contact with may not adequately cover the options in the given area. The second option for patients is to seek public information about and/or ratings for a physician available on, e.g., the internet. Such ratings, however, are sparse as medical history is often treated as personal, confidential information. Public ratings also suffer from the problem of trustworthiness, as the likelihood of inaccuracies is higher.

II. LITERATURE REVIEW

Background Study

There has been considerable research into privacy preserving recommendation systems. Originally, privacy was achieved in recommendation systems by giving user information to a trusted third party, who then performs the necessary calculations with other trusted agents. One problem with this early approach is that, in addition to privacy, in order to be useful, recommendation systems must be robust against misbehaving users. One common way misbehaving users may attempt to influence the rating of a specific physician is known as “shilling attacks.”

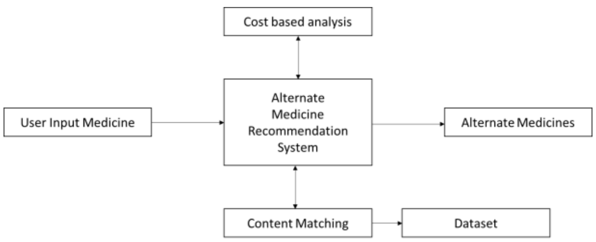
Shilling attacks are said to occur when a user attempts to sabotage a competitor in order to make themselves look better. Lam and Reidl [Lam. and Riedl 2004] describe the attacks and discuss how they can affect the recommender system. Specifically, the authors consider various attack motivations (e.g., increasing/decreasing, the rating of an item and hindering the credibility of the recommendation system as a whole) and their effect on recommendation systems.

Importantly, they note that while observing sharp changes in scores is an obvious way to detect (some) shilling attacks, non-trivial attacks against the system could potentially succeed. Detecting such attacks is proposed as a future area of research. Chirita, Nejdl, and Zamfir [Chirita et al. 2005] provide further insight into shilling attacks and outline a detection algorithm which depends on the distribution of scores that each user has made so far. The algorithm proved to be quite robust, providing not many false positives while catching many of the shilling attacks. In this work, on the other hand, instead of trying to detect system abuse, we concentrate on abuse prevention.

While in shilling attacks competing physicians attempt to sabotage each other, "bad mouthing" [Bankovic et al. 2011] is said to occur when a (potentially offended) patient attempts to lower the score of a physician. "Boosting" (or "ballot stuffing") is said to occur if, instead of lowering a score, the patients collude to increase a rating [Dellarocas 2000; Srivatsa et al. 2005].

III. PROPOSED METHODOLOGY

Proposed System Architecture



Dataset Generation

Medicine dataset needs to be generated for multiple medicines having same ratio which can be provided as an alternative to each other. Also for cost analysis it is necessary to know the market cost for such medicines. This dataset will be created for a total of 50 medicines.

Dataset Preprocessing

As dataset gets generated it is necessary to preprocess it for any null values if provided and the data should be cleaned and stored into Database for further processing.

Data Clustering based on contents and costing

Data clustering needs to be done for grouping similar medicines based on their contents and also it is required for cost based analysis as well.

Medicine classification and recommendation

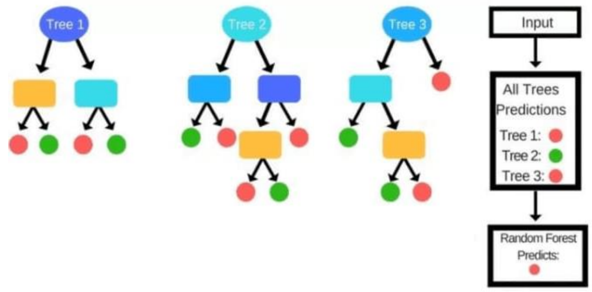
At last ones the user inputs some medicine there is a requirement of finding alternate medicines for users which can only be done using classification of input medicines using some classification algorithm.

IV. ALGORITHMS

1. Random Forest Algorithm

Random Forest is a supervised learning algorithm. Like you can already see from it's name, it creates a forest and makes it somehow random. The „forest" it builds, is an ensemble of Decision Trees, most of the time trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.

To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

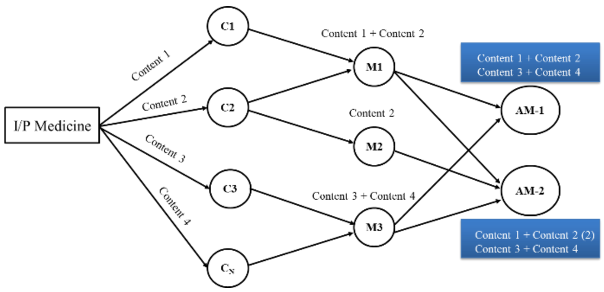


One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems. I will talk about random forest in classification, since classification is sometimes considered the building block of machine learning.

Random Forest has nearly the same hyperparameters as a decision tree or a bagging classifier. Fortunately, you don't have to combine a decision tree with a bagging classifier and can just easily use the classifier-class of Random Forest. Like I already said, with Random Forest, you can also deal with Regression tasks by using the Random Forest regressor. Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Therefore, in Random Forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make trees more random, by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does).

Working Example



V. CONCLUSION

This project explores the strong application of data mining in the field of medical recommendation systems. In this project we have used a dataset over 100+ medicines from different companies and brands to do recommendation based on content of medicine and then filter it based on rating and cost based analysis. Through experimental results we have found that more than 95% of the medicines have a lower cost based alternative available with a higher rating. We have used random forest algorithm for classification of medicines based on the costing and rating of the medicines provided by users. Also we have made a comparison of three different algorithms k-NN decision tree and random forest algorithm for classification of medicines in particular alternatives.

VI. REFERENCES

[1]. MacLennan A, Myers S, Taylor A: The continuing use of complementary and alternative medicine in South Australia: costs and beliefs in 2004. Med J Aust. 2006, 184: 5

[2]. Xue CCL, Zhang AL, Lin V, Da Costa C, Story DF: Complementary and alternative medicine use in Australia: a national population-based survey. J Altern Complement Med. 2007, 13: 643-650. 10.1089/acm.2006.6355.

[3]. Braun L: Integrative Pharmacy: threat or fantasy?. J Compl Med. 2009, 8: 41-42.

[4]. Semple SJ, Hotham E, Rao D, Martin K, Smith CA, Bloustien GF: Community pharmacists in Australia: barriers to information provision on complementary and alternative medicines. Pharm World Sci. 2006, 28: 366-373.

[5]. Braun L, Tiralongo E, Wilkinson J, Spitzer O, Bailey M, Poole S, Dooley M: Perceptions, use and attitudes of pharmacy customers on complementary medicines and pharmacy practice. BMC Complement Altern Med. 2010, 10: 38-10.1186/1472-6882-10-38.