TRANSFERRING KNOWLEDGE for FEATURE EXTRACTION for DRUG TOXICITY PREDICTION USING UTILITY COMBINATIONS

M.S.Danessh [#1], S.Vasanth [*2]

[#] *CSE Department, K.S.Rangasamy College of Technology*
Tiruchengode, India
[1] visitdanesh@gmail.com
[*] *K.S.Rangasamy College of Technology*
Tiruchengode, India
[2] vasanthcs51@gmail.com

Abstract - **Utilizing the available data to increase the predictive performance on newly designed task is the main problem in data mining. In this paper, the main goal is to transfer data from source to target (client to server). The source (client) information is transferred to target (server) for drug toxicity prediction. For effective transfer of knowledge, effective data representation is needed for new modeling task. Feature extraction has been used in data mining for representation of subsequent modeling. Most common method used for feature extraction is Principle Component Analysis (PCA). The technique used in this paper is sparse coding, for assigning feature values based on cluster groups. Sparse coding can identify the higher order feature from the raw representation. Sparse coding is suitable for subsequent analysis which includes subspace clustering. Many drug toxic reactions are not discovered during limited pre-marketing clinical trials instead, it only observed after long term post-marketing surveillance of drug usage. The detection of adverse drug reactions is an important topic of research for the medicinal industry. Recently, adverse events of large numbers and the development of data mining technology have motivated the development of statistical and data mining methods for the detection of DTRs. The proposed two algorithms, namely utility pattern growth (UP-Growth) and UP-Growth+, for mining the utility item sets with a set of effective strategies for pruning candidate item sets. The information of utility item sets is maintained in a tree-based data structure named utility pattern tree (UP-Tree) such that candidate item sets can be generated efficiently with only two scans of database. The UP-Growth+ and UP-Growth performance is compared with the state-of-the-art algorithms on different types of both synthetic and real data sets.**

Keywords – **Knowledge transfer, feature extraction, utility pattern growth.**

I. INTRODUCTION

Effectively using readily available auxiliary data to reform predictive performance on new modeling tasks is an important problem in data mining. Most commonly used feature extraction methods is Principle Component Analysis (PCA) [3]. PCA methods can perform feature extraction for knowledge transfer tasks. The application of PCA-based methods for knowledge transfer has two various reasons [1]. One is for different distributions of source data and target data can spoof the direction of principle components. Other one is for high dimensional data; the data is clustered only in subspaces rather than full space [5]. PCA can not notify the representation of the data. Towards the end goal of the effective data representation, sparse coding is used. Sparse coding is used for identifying a group of higher order features of data from the raw data representations. The disadvantage of the sparse coding is that the distribution distance has some problem for knowledge transfer. The proposed method with synthetic and real data experiments with application to drug toxicity prediction is evaluated. For example, the finding drug adverse (FDA) currently adopts a data mining algorithm called Multi-item Gamma Poisson Shrinker for detecting potential signals from its original reports. Other important signal detection strategy is known as the Bayesian Confidence Propagation Neural Network that has been used by the Uppsala Monitoring Center in routine with its World Health Organization database. As electronic patient records become more and more easily accessible in various health organizations such as hospitals and medical centers, they provide a new source of information that has great potential to generate ADR signals much earlier [4]. Note that each patient case can be considered as an event sequence where events such as drug(tablet) prescription, event of symptom and laboratory test occur at alternate times.

The ultimate goal of drug utilization research must be to assess whether drug therapy is rational or not. Towards the goal, methods for auditing drug therapy towards rationality are required [3]. The previous work did not allow detailed comparisons of the drug utilization data obtained from different users because the source and form of the information varied between them.

II. Sparse Coding for Feature Extraction in Knowledge Transfer

Sparse coding is used for transfer learning that can capture higher level feature of data to allow knowledge transfer [1]. The shared features can build the regression models for prophet the missed values and also find the lower dimensional shape and allowing the data for knowledge transfer [4].

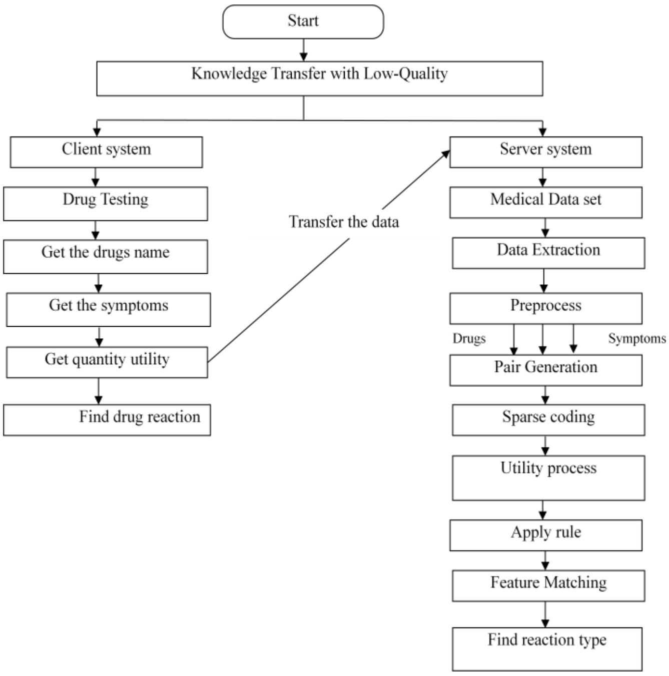

Fig. 1 Block Diagram for Knowledge Transfer for Drug Toxicity Prediction

The approach is that, imputation and learning the embedding can be performed individually, also the structure tells the missing values so that the latent shape can be studied. For example, the finding drug adverse (FDA) currently adopts a data mining algorithm. That is, there are 50 in number. Around 50 drugs are picked randomly from the drug list (FDA) and constitute a class.

Figure1 shows the knowledge transfer with low quality from client and server. The client (user) can get the approval for drug and also view the side effects. That is, the client transfers the knowledge (information) to server by giving the requirements mentioned (drug name, quantity, symptom). The server will get the information from client and extract the feature from sparse data. Then, calculate the entropy and display the results.

III. Transfer Learning

Any number of learning algorithms has been developed for transferring knowledge. One of the most used approaches is model-based approach where different distributions are equaled in a model. Another approach to develop the model is transductive transfer learning which refers local structure of the unlabeled data [5]. The model selection and selecting features method can generalizes the distributions. The feature selection and feature generation is compared to discover the new features for knowledge transfer and also in regularization framework.

The transfer learning has three different settings which has four cases [5],

- instance-based transfer learning
- feature-representation-transfer
- parameter-transfer
- relational-knowledge-transfer

A. Instance-based transfer learning

Some parts of data in the source domain should be reused for learning in target domain.

B. Feature-representation-transfer

The knowledge can be used to transfer across domains into learned feature representations.

C. Parameter-transfer

The transferred knowledge can be encoded into the shared parameters

D. Relational-knowledge-transfer

The knowledge can be transferred is the relationship among the data.

IV. Domain Adaptation

Domain adaptation allows knowledge transfer from source domain and transferred to related target domain. To overcome such representation transfer component analysis (TCA) is used [5]. The common feature extraction approach where TCA, tries to learn set of some transfer components underlying both source and target domain.

V. Feature Extraction Using Sparse Coding

One of the advantages of sparse coding is that learning higher order representation of data from the given low level representation. Other way of viewing the sparse coding which offers more insight for geometric perspective. Sparse coding can performs subspace clustering [1]. Sparse coding can be used to identify the subspace clusters [4]. It is useful for knowledge transfer in the same sense the clustering based transfer learning can identify the shared cluster structure of data with the goal data. Sparse coding is one of the class of unsupervised methods for learning sets of over-complete basis to represent data effectively.

$$x = \sum_{i=1}^{k} a_i \varphi_i$$

The main aim of sparse coding is to find the basis vector φ_i where input vector x as linear combination of basis vector.

VI. Incorporating Target Data Label Information

A common data mining or knowledge discovery task focuses on classification. The data from ground truth label information is expensive and time consuming to obtain, only a small amount of label information is to be obtained. The sparse coding with distribution distance will tries to approximate the data well. The incorporation of class-based distribution distance should be based on some theoretical results for knowledge transfer [1]. The theoretical upper bounds on target error take the form of source error with distribution distance based on the marginal distributions.

VII. Up Growth

The proposed two algorithms, namely utility pattern growth (UP-Growth) and UP-Growth+, for mining the utility item sets with a set of effective strategies for pruning candidate item sets. The information of utility item sets is maintained in a tree-based data structure named utility pattern tree (UP-Tree) such that candidate item sets can be generated efficiently with only two scans of database. The UP-Growth+ and UP-Growth performance is compared with the state-of-the-art algorithms on different types of both synthetic and real data sets. Experimental results shows the proposed algorithms, the UP-Growth+ not only minimize the number of candidates effectively but also outperform other algorithms substantially in terms of runtime, particularly when databases contain number of long transactions.

VIII. Conclusion

Knowledge transfer has attracted from many learning and data mining algorithms. Knowledge transfer focuses in different direction and deals with preprocessing techniques. Data without ground truth information from different distribution to aid in knowledge discovery. After investigated the sparse coding, it is used for identifying a group of higher order features of data from the raw data representations. The disadvantage of the sparse coding is that the distribution distance has some problem for knowledge transfer. The proposed method with synthetic and real data experiments with application to drug toxicity prediction is evaluated. For example, the finding drug adverse (FDA). The proposed two algorithms, namely utility pattern growth (UP-Growth) and UP-Growth+, for mining the utility item sets with a set of effective strategies for pruning candidate item sets.

References

[1] *Brian Quanz, Jun (Luke) Huan and Meenakshi Mishra (2012), "Knowledge Transfer with Low-Quality Data: A Feature Extraction Issue", IEEE Knowledge and Data Engineering, Vol. 24, no. 10, pp. 1789- 1802.*

[2] *G. Xue, Q. Yang, W. Dai, X. Ling and Y. Yu (2008), "Spectral Domain-Transfer Learning" Knowledge Discovery and Data Mining, pp. 488 - 496.*

[3] *Hao Ying and Yanqing Ji (2013), "A Method for Mining Infrequent Causal Associations and Its Application in Finding Adverse Drug Reaction Signal Pairs", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, no. 4, pp. 721- 733.*

[4] *Ivor Kwok, James Tsang, Qiang Yang and Sinno Jialin Pan (2012), "Domain Adaptation via Transfer Component Analysis", IEEE Knowledge and Data Engineering, Vol. 22, no. 2, pp. 199- 210.*

[5] *Qiang Yang and Sinno Jialin Pan (2010), "A Transfer Learning Survey", IEEE Transactions on Data and Knowledge Engineering, Vol. 22, no. 10, pp. 1345- 1359.*