International Journal of Scientific Research in Computer Science, Engineering and Information Technology

© 2018 IJSRCSEIT | Volume 3 | Issue 8 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT183813

Drug Prediction System Using Data Mining Techniques - A Survey

V. Jagadeesan[1*], Dr. K. Palanivel[2]

[1*]Research Scholar, Department of Computer Science, A.V.C. College (Autonomous), Mayiladuthurai, India

[2]Research Supervisor, Department of Computer Science, A.V.C. College(Autonomous), Mayiladuthurai, India

ABSTRACT

The thriving Medical applications of Data mining in the fields of Medicine and Public health has led to the popularity of its use in Knowledge Discovery in Databases (KDD). Data mining has revealed novel Biomedical and Healthcare acquaintances for Clinical decision making that has great potential to improve the treatment quality of hospitals and increase the survival rate of patients. Drug Prediction is one of the applications where data mining tools are establishing the successful results. Data mining intends to endow with a systematic survey of current techniques of Knowledge discovery in Databases using Data mining techniques that are in use in today's Medical research. To enable the drug retrieval and the breakthrough of hidden retrieval patterns from related databases, a study is made. Also, the use of data mining to discover such relationships as those between Supervised and Unsupervised are presented. This paper summarizes various Machine learning algorithms based on various Data mining techniques in learning strategies. It has also been targeted on contemporary research being done the usage of the Data mining strategies to beautify the retrieval manner. This research paper offers destiny developments of modern-day strategies of KDD, using data mining equipment for medicinal drug industry. It also confers huge troubles and demanding situations related to information mining and medication area. The research discovered a developing quantity of records mining packages, such as evaluation of drugs names for higher fitness policy-making, detection of accurate effects with outbreaks and preventable from misclassified drug names.

Keywords : Drug Query System, Knowledge Discovery in Database, Supervised, Unsupervised and Semi-supervised.

I. INTRODUCTION

Several precautions should be taken in using pharmaceutical drugs, for both healthcare professionals, who prescribe and administer drugs, and for drug consumers. Factors such as interactions among the prescribed drugs, interactions with the patient's current medication, side effects to be avoided, and contraindications, need to be carefully considered. Additionally, the presence of some drug properties, such as side effects and effectiveness, depends on characteristics of patients, such as age, gender, lifestyles, and genetic profiles. Having to consider all these complicated factors can be a huge burden to professionals and drug consumers. There is one old proverb which states that Good medicine takes bitter in mouth. From these we can understand that medicine plays a vital and more important role in our life. The medical mafia, people who were affected by infections or disease suffers a lot with infection as well as medicine inhabitation. People who are taking medicine should undergo numerous precautions and effective measures. In this present system, if a person is affected by disease, he will consult doctor and the doctor will prescribe some tablets. These tablets may contain some side effects because we take lot of medicines. In some medicines they will mention regarding this. i.e. For example – in medicine itself they mention that if the medicine is consumed for a longer period of time, these side effects will happen. Many people are unaware of it. Some people have not to take the medicine itself. But for some disease let us say diabetes or hair fall or thyroid , we may consume many tablets which may causes side effects, but the doctors will analysis regarding the medicine should be given to patients or not. But still some doctors do not analysis the side effects correctly, so we cannot say full proof or idle for the existing system. Basically we approach a doctor and we go by doctor opinion. Mostly people will not think to get the second opinion. For example diseases like fever, cold, headache people do not get the second opinion. But for some disease we should get second opinion but that too if the consulting doctor was famous, from him it is questionable to consult for second doctor and it will be complicated. So purely we had to rely on doctors. In present system they mention that one or more tablets may have some side effects in which is analyzed by patients profile such as age, sex, genetic information and life style. It states that particular medicine is taken by patients for long period and these side effects may occur. It is one of major core concept in existing system.

Many efforts have been made in providing drug related information to physicians as well as laypersons through the means of publishing drug information resources – printed, as well as in digital form (through CDs, free information sites on the internet, etc.), web interfaces where one can submit the queries for them to be answered by experts in the respective fields or one can visit drug information centers (DICs) to get his or her questions answered. DICs are primarily operated by medical schools or respective department/ministry of the government. These centers answer drug and drug therapy related questions both from professionals and general public via phone calls, email or fax. Additionally, numerous online drug information sites provide forums for discussing patient-specific problems/queries along with detailed factual information about drugs. These sites report tens of millions of unique visitors per month. Many printed drug reference sources are easily available starting from handbooks to detailed guides for prescription and non-prescription drugs. These drug information sources however are either unknown or inaccessible to the semi-urban or rural population. Also, many of them such as the DICs and the printed books require delayed processing wherein one has to manually search through the available material for the required information. This information is also accessible via the internet but the internet is yet to become ubiquitous, particularly in developing countries. Drug related information can be useful to the general public in emergency situations when one has many medicines lying in his or her medicine-box but does not know which one to use in that situation. In such circumstances, the patient or anybody else in the vicinity of the patient would want to know various aspects of a given drug like its usage, dosage, side-effects, directions on how to take the drug, etc. In a developing country like India where doctor-to-patient ratio is very low availability of such information becomes extremely important because then the situation can be tackled based on evidence without requiring to contact a physician who is generally not easily reachable. Some other situations in which such information can be useful are – when one wants to know the adverse interactions between two or more drugs that he or she is taking or the precautions to be taken while on a medication or whether a drug is contraindicated in case of old age, pregnancy, etc. Availability of drug related information instantaneously in the absence of accessibility to the earlier mentioned resources makes it an important domain of interest. The basic layout of drug query system is shown in figure 1.

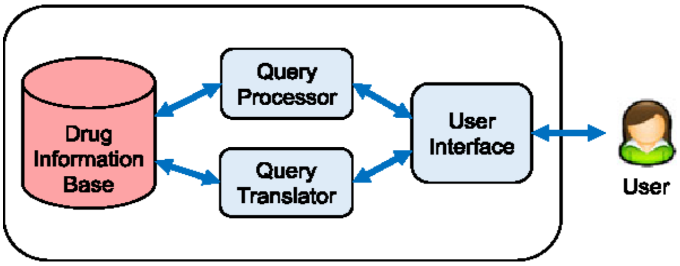

Figure 1 : Drug Query Processing

The Drug query system has drug information database and query processor to get the supervised results from database based user's search query. Various data mining techniques are needed to extract the query results from database.

Data mining techniques

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. An interdisciplinary subfield of computer science, it is an essential process wherein intelligent methods are applied to extract data patterns the overall goal of which is to extract information from a data set, and transform it into an understandable structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Health organizations today are capable of generating and collecting a large amount of data. This increase in data volume automatically requires the data to be retrieved when needed. With the use of data mining techniques is possible to extract the knowledge and determine interesting and useful patterns. The knowledge gained in this way can be used in the proper order to improve work efficiency and enhance the quality of decision making. Above the foregoing is a great need for new generation of theories and computational tools to help people with extracting useful information from the growing volume of digital data. Information technologies are implemented increasingly often in healthcare organizations to meet the needs of physicians in their daily decision making. Computer systems used in data mining can be very useful to control human limitations such as subjectivity and error due to fatigue and to provide guidance to decision-making processes. There are several major *data mining techniques* have been developing and using in data mining projects recently including *association*, *classification*, *clustering*, *prediction, sequential patterns* and *decision tree*.

Association

Association is one of the best-known data mining techniques. In association, a pattern is discovered based on a relationship between items in the same transaction. That's the reason why association technique is also known as relation technique. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together. Retailers are using association technique to research customer's buying habits. Based on historical sale data, retailers might find out that customers always buy crisps when they buy beers, and, therefore, they can put beers and crisps next to each other to save time for the customer and increase sales.

Classification

Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network, and statistics. In classification, we develop the software that can learn how to classify the data items into groups. For example, we can apply classification in the application that "given all records of employees who left the company; predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups.

Clustering

Clustering is a data mining technique that makes a meaningful or useful cluster of objects which have similar characteristics using the automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. To make the concept clearer, we can take book management in the library as an example. In a library, there is a wide range of books on various topics available. The challenge is how to keep those books in a way that readers can take several books on a particular topic without hassle. By using the clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in that topic, they would only have to go to that shelf instead of looking for the entire library.

Prediction

The prediction, as its name implied, is one of a data mining techniques that discover the relationship between independent variables and relationship between dependent and independent variables. For instance, the prediction analysis technique can be used in the sale to predict profit for the future if we consider the sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

Sequential Patterns

Sequential patterns analysis is one of data mining technique that seeks to discover or identify similar patterns, regular events or trends in transaction data over a business period.In sales, with historical transaction data, businesses can identify a set of items that customers buy together different times in a year. Then businesses can use this information to recommend customers buy it with better deals based on their purchasing frequency in the past.

Decision trees

The decision tree is one of the most commonly used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it.

This paper is organized as, the Section I consists of brief introduction of drug query system, data mining and techniques in data mining. Section II elaborates the related work. Section III provides the details about existing methodologies. Section IV explains the proposed work with provides comparative analysis in Section V. In Section VI, conclude the overall survey paper.

II. RELATED WORK

Emmanuel Bresso, et.al,…[1] proposed drug annotations are collected from SIDER and DrugBank databases. Terms describing individual side effects reported in SIDER are clustered with a semantic similarity measure into term clusters (TCs). Maximal frequent item-sets are extracted from the resulting drug x TC binary table, leading to the identification of what we call side-effect profiles (SEPs). A SEP is defined as the longest combination of TCs which are shared by a significant number of drugs. Frequent SEPs are explored on the basis of integrated drug and target descriptors using two machine learning methods: decision-trees and inductive-logic programming. Although both methods yield explicit models, inductive-logic programming method performs relational learning and is able to exploit not only drug properties but also background knowledge. Learning efficiency is evaluated by cross-validation and direct testing with new molecules. Comparison of the two machine-learning methods shows that the inductive-logic-programming method displays a greater sensitivity than decision trees and successfully exploits background knowledge such as functional annotations and pathways of drug targets, thereby producing rich and expressive rules. And also proposed an approach based on the hypothesis that drugs sharing side effects could be indicated for the same disease. Drug side-effect associations and drug disease relationships were used to develop a systematic drug repositioning method and to suggest, for instance, an anti-diabetic effect for drugs causing porphyria.

Assaf Gottlieb, et.al,…[2] presented an approach for predicting novel associations between drugs and diseases that can operate on both drugs with approved indications and on novel molecules with no indication information. Given a query association, we measure the similarity of the pertaining drug and disease to drug–disease pairs that are known to be associated, and rank the accumulative evidence for association using a logistic regression scheme. The prediction process is aided by a comprehensive drug–disease association data set that have compiled and a collection of novel drug–drug similarity measures. Importantly, show the potential utility of approach also in a personalized medicine setting, in which a disease name is replaced by a gene expression signature; and consequently, disease–disease similarity is measured via the similarity of the corresponding signatures. And designed a novel algorithm for PREdicting Drug IndiCaTions (PREDICT). Given a gold standard set of drug–disease associations (known associations), the algorithm ranks additional drug–disease associations based on their similarity to the known associations. The algorithm works in three phases (Figure 1): (i) construction of drug–drug and disease– disease similarity measures; (ii) exploiting these similarity measures to construct classification features and subsequent learning of a classification rule that distinguishes true from false drug–disease associations; and (iii) application of the classifier to predict new associations.

David, et.al,…[3] offered a generic text search, a local BLAST search (SeqSearch), a higher level Boolean text search (TextQuery), a chemical structure search utility (ChemQuery) and a relational data extraction tool (Data Extractor). Each of these search utilities has a number of useful bioinformatics or cheminformatic applications, many of which were described in the first DrugBank publication. For the latest release of DrugBank, have added a number of improvements to both the generic text search and ChemQuery. In particular, the generic text search has been enhanced so that users now have the option of clicking on check boxes to limit their search to a drug's common name, its synonyms/ brand names or all text fields. Because the vast majority of queries to DrugBank are related to drug names/synonyms, the default query always has these two boxes checked off. Users wishing to search through the other 100+ data fields in DrugBank can select the 'all text fields' box. This change has also substantially improved the query response times for most DrugBank text searches. Because the spelling of many drug names, chemical compound names and protein names is often difficult or non-intuitive, DrugBank now supports an 'intelligent' text search, where alternative spellings to misspelled or incompletely entered names are automatically provided. In addition to this change, the results from text queries have also been enhanced so that the standard tabular output (primary accession number, generic drug name, chemical formula and molecular weight) is supplemented with the query word highlighted in the selected DrugCard field(s) from which it was retrieved.

Joanne Bowes, et.al,…[4] evaluated potential side effects of drugs is important in rational drug design and development, as well as successful marketing. Binding of drugs to their on- and off-targets modifies the functions of these targets and therefore is believed to account for their efficacies as well as side effects. Traditionally, properties of a drug such as binding fingerprint and chemical structure are evaluated to anticipate side effects. Moreover, in vitro assays or phenotypic tests in model organisms may not be able to capture the same spectrum of side effects in human. Recently, an increasingly accepted view is that integrating biological networks would provide unique insights into understanding disease mechanisms and identifying novel drug targets. Network-based methods have been explored and successfully applied in finding disease-associated genes and inferring underlying molecular mechanisms. Similarly, phenotypic responses to drugs can be better rationalized by considering their overall effects in the context of molecular networks. Previous studies have shown that drugs with shared targets or those that are close in the interact the network often share similar side effects. Also, similar side effect profiles have been used to predict drug-target interactions for potential drug repositioning.

Xiujuan Wang, et.al,…[5] used the high-capacity in vitro pharmacology panels is one aspect of a broader trend in focusing on drug safety much earlier in the drug discovery process, with the aim of reducing the high rate of attrition. Selecting the minimal number of targets, and deciding which targets to include, in an in vitro pharmacological profiling assay is an exercise in judgment and experience, and also depends on budgetary and technical constraints. In summary, in vitro pharmacological profiling is a valuable tool that can allow the early identification of off-target pharmacological interactions that could cause safety liabilities in the clinic, and this early identification of safety liabilities could improve decision-making by discovery project teams. The use of the minimal panel of targets recommended in this article might help to reduce safety-related attrition of drug molecules during drug discovery and development. Further precompetitive knowledge management of this data could lead to the development of in silicotools that more accurately predict pharmacological activity and integration of these data with robust in vivo models could enable efficient and cost-effective early decision-making based on accurate predictions of the exposures at which a safety liability may be expected in the human population. We hope that this article is a first step towards establishing a broad initiative to work closely on improving drug safety from early stages of drug discovery through to clinical development and at the post-marketing stage.

III. EXISTING METHODOLOGIES

One of the critical stages in drug development is the identification of potential side effects for promising drug leads. Large scale clinical experiments aimed at discovering such side effects are very costly and may miss subtle or rare side effects. Identification of underlying mechanisms behind drugs side effects is of extreme interest and importance in drugs discovery today. Therefore machine learning methodology, linking such different multi features aspects and able to make predictions, are crucial for understanding side effects.The existing methodologies use the supervised and unsupervised learning systems in drug query system.

Supervised learning algorithm:

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples.[2] In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

The steps are defined in supervised learning can be listed as follows:

- Determine the type of training examples. Before doing anything else, the user should decide what kind of data is to be used as a training set. In case of handwriting analysis, for example, this might be a single handwritten character, an entire handwritten word, or an entire line of handwriting.
- Gather a training set. The training set needs to be representative of the real-world use of the function. Thus, a set of input objects is gathered and corresponding outputs are also gathered, either from human experts or from measurements.
- Determine the input feature representation of the learned function. The accuracy of the learned function depends strongly on how the input object is represented. Typically, the input object is transformed into a feature vector, which contains a number of features that are descriptive of the object. The number of features should not be too large, because of the curse of dimensionality; but should contain enough information to accurately predict the output.
- Determine the structure of the learned function and corresponding learning algorithm. For example, the engineer may choose to use support vector machines or decision trees.
- Complete the design. Run the learning algorithm on the gathered training set. Some supervised learning algorithms require the user to determine certain control parameters. These parameters may be adjusted by optimizing performance on a subset (called a validation set) of the training set, or via cross-validation.
- Evaluate the accuracy of the learned function. After parameter adjustment and learning, the performance of the resulting function should be measured on a test set that is separate from the training set.

The mostly used supervised learning algorithms are:

- Support Vector Machines
- Linear Regression
- Logistic Regression
- Naive Bayes
- Linear Discriminant Analysis
- Decision trees
- K-Nearest Neighbor algorithm

The supervised learning algorithm provided computational complexity because of large number of datasets for trained and also support only case sensitive query.

Unsupervised learning algorithm:

Unsupervised learning is a branch of machine learning that learns from test data that has not been labeled, classified or categorized. Instead of responding to feedback, unsupervised learning identifies commonalities in the data and reacts based on the presence or absence of such commonalities in each new piece of data. Alternatives include supervised learning and reinforcement learning.A central application of unsupervised learning is in the field of density estimation in statistics, though unsupervised learning encompasses many other domains involving summarizing and explaining data features.Compared to supervised learning where training data is labeled with the appropriate classifications, models using unsupervised learning must learn relationships between elements in a data set and classify the raw data without "help." This hunt for relationships can take many different algorithmic forms, but all models have the same goal of mimicking human logic by searching for indirect hidden structures, patterns or features to analyze new data.

The mostly used unsupervised learning algorithms are:

- Principal component analysis
- Independent component analysis
- Non-negative matrix factorization
- Singular value decomposition

The classical example of unsupervised learning in the study of neural networks is Donald Hebb's principle, that is, neurons that fire together wire together. In Hebbian learning, the connection is reinforced irrespective of an error, but is exclusively a function of the coincidence between action potentials between the two neurons. A similar version that modifies synaptic weights takes into account the time between the action potentials (spike-timing-dependent plasticity or STDP). Hebbian Learning has been hypothesized to underlie a range of cognitive functions, such as pattern recognition and experiential learning.One of the statistical approaches for unsupervised learning is the method of moments. In the method of moments, the unknown parameters (of interest) in the model are related to the moments of one or more random variables, and thus, these unknown parameters can be estimated given the moments. The moments are usually estimated from samples empirically. The basic moments are first and second order moments. For a random vector, the first order moment is the mean vector, and the second order moment is the co-variancematrix (when the mean is zero). Higher order moments are usually represented using tensors which are the generalization of matrices to higher orders as multi-dimensional arrays. In particular, the method of moments is shown to be effective in learning the parameters of latent variable models.Latent variable models are statistical models where in addition to the observed variables, a set of latent variables also exist which are not observed. A highly practical example of latent variable models in machine learning is the topic modeling which is a statistical model for generating the words (observed variables) in the document based on the topic (latent variable) of the document. In the topic modeling, the words in the document are generated according to different statistical parameters when the topic of the document is changed. It is shown that method of moments (tensor decomposition techniques) consistently recovers the parameters of a large class of latent variable models under some assumptions. The major disadvantage of the unsupervised algorithm can be provided irrelevant results in drug query system.

IV. OUTCOME OF SURVEY

Table 1. provides comparative analysis of reviewed papers.

Author names	Methodology	Performance	Advantages	Disadvantages
Emmanuel Bresso,et.al,…[1]	Inductive Logic Programming (ILP)	Uses relational data as input and has been successfully applied to various areas including bioinformatics	Provide first-order logic rules	Maximal number of negative examples
Assaf Gottlieb, et.al,…[2]	PREdicting Drug IndiCaTions (PREDICT)	Ranks additional drug–disease associations based on their similarity to the	Eliminate the redundant drugs in query	Computational complexity is high

	framework	known associations.	system	
David, et.al,…[3]	Drug Bank Framework	Annotated resource that combines detailed drug data with comprehensive drug	Response to numerous user requests	Need manual annotation efforts for drug datasets
Joanne Bowes, et.al,…[4]	Vitro pharmacological profiling	Targets include representatives from the major neurotransmitter classes	Can be tested cost-effectively in vitro on human targets	Lack of selective query from databases
Xiujuan Wang, et.al,…[5]	Generalized linear regression analysis	Contributing to the incidence of side effects, performed a series of generalized linear regressions based on negative binomial distribution	Obtained the disease-associated genes	Large number continuous datasets are needed

Table 1: Comparative Analysis

Extracted results used in learning to analyze the performance of the system in terms of accuracy are given in Table 2. Existing system sometimes make classification more complicated. It is required to increase the irrelevant results. Reduce accuracy in terms of case sensitive results, but semi-supervised approach can act as perfect approach during training part as well as testing part. For quantitative analysis, performance of learning algorithms are surveyed and shown in table 2.

Category	Method	Performance
Supervised learning	Naives Bayes algorithm	Provide probability based results
	K-NN algorithm	High level response time
	Decision tree algorithm	Tree structure become complex
	Linear Regression algorithm	Continuous data needed
Unsupervised Learning	Principal component analysis	Dimensionality is high
	Linear discriminative Analysis	Error Rate can be occurred
	Singular value decomposition	Irrelevant results are extracted
Semi-supervised learning	Semi-supervised Support Vector Machine (SSVM)	Improved accuracy with relevant results extracted

Table 2 : Performance Table

V. PROPOSED FRAMEWORK

From the above sections, we analyzed the issues in supervised and unsupervised learning algorithms. To overcome the problems, we can propose to do semi-supervised learning in drug query system.Semi-supervised learning is a class of machine learning tasks and techniques that also make use of unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data.). Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy. The acquisition of labeled data for a learning problem often requires a skilled human agent or a physical experiment. The cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is relatively inexpensive. In such situations, semi-supervised learning can be of great practical value. Semi-supervised learning is also of theoretical interest in machine learning and as a model for human learning. The proposed system makes full use of the unlabeled instance. This method goes with this assumption that the change of all the labels should be smooth on the graph.Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy over unsupervised learning (where no data is labeled), but without the time and costs needed for supervised learning (where all data is labeled). The acquisition of labeled data for a learning problem often requires a skilled human agent or a physical experiment. The cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is relatively inexpensive. In such situations, semi-supervised learning can be of great practical value. Semi-supervised learning is also of theoretical interest in machine learning and as a model for human learning. The major assumption is query. Queries which are close to each other are more likely to share a label. This is also generally assumed in supervised learning and yields a preference for geometrically simple decision boundaries. In the case of semi-supervised learning, the smoothness assumption additionally yields a preference for decision boundaries in low-density regions, so that there are fewer points close to each other but in different classes.

Semi-supervised learning can be categorized into two slightly different settings, denoted transductive and inductive learning.

1. Transductive learning: It concerns the problem of predicting the labels of the unlabeled examples, given in advance, by taking both labeled and unlabeled data together into account to train a classifier.
2. Inductive learning: It considers the given labeled and unlabeled data as the training examples, and its objective is to predict unseen data.

   The main purpose of metric learning in specific problem is to obtain an appropriate distance /similarity function. User can input the drug name or side effects to classify the results with improved accuracy. We can implement semi-supervised classification algorithm to get the results based on side effects. User can get the results of drug with improved side effects.

VI.CONCLUSION

Using data mining technology for disease prediction and diagnosis has become the focus of attention. Data mining technology provides an important means for extracting valuable medical rules hidden in medical data and acts as an important role in disease prediction and clinical diagnosis. In the current study, have demonstrated, using a large sample of patients hospitalized with classification. As discussed in above survey, prediction of drug query system has several approaches such as supervised, unsupervised and semi-supervised approaches. Based on this supervised approach only implemented in trained datasets with proper format. Unsupervised algorithm may provide results with limited accuracy. We find that semi-supervised learning produced the better result in comparison with supervised method in drug retrieval system.

VII. REFERENCES

[1]. E. Bressoet al., "Integrative relational machine-learning for understanding drug side-effect profiles", BMC Bioinf., vol. 14, Jun. 2013, Art.no. 207.

[2]. T. Liu and R. B. Altman, "Relating essential proteins to drug side effects using canonical component analysis: A structure-based approach"J. Chem. Inf. Model., vol. 55, no. 7, pp. 1483_1494, 2015.

[3]. D. S.Wishartet al., "DrugBank: A knowledgebase for drugs, drug actions and drug targets", Nucl. Acids Res., vol. 36, pp. D901_D906, Nov. 2008.

[4]. J. Bowes et al., "Reducing safety-related drug attrition: The use of in vitro pharmacological profiling", Nature Rev. Drug Discovery, vol. 11, no. 12, pp. 909_922, 2012.

[5]. X. Wang, B. Thijssen, and H. Yu, "Target essentiality and centrality characterize drug side effects", PLoSComput.Biol., vol. 9, no. 7, p. e1003119, 2013.

[6]. M. Duran-Frigola and P. Aloy, "Analysis of chemical and biological features yields mechanistic insights into drug side effects", Chem. Biol., vol. 20, no. 4, pp. 594_603, 2013.

[7]. T. Liu and R. B. Altman, "Relating essential proteins to drug side effects using canonical component analysis: A structure-based approach"J. Chem. Inf. Model., vol. 55, no. 7, pp. 1483_1494, 2015.

[8]. S. Jamal, S. Goyal, A. Shanker, and A. Grover, "Predicting neurological adverse drug reactions based on biological, chemical and phenotypic properties of drugs using machine learning models"Sci. Rep., vol. 7, Apr. 2017, Art. no. 872.

[9]. J. Scheiberet al., "Mapping adverse drug reactions in chemical space",J. Med. Chem., vol. 52, no. 9, pp. 3103_3107, 2009.

[10]. Y. Yamanishi, M. Kotera, M. Kanehisa, and S. Goto, "Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework" Bioinformatics, vol. 26, no. 12, pp. i246_i254, 2010

[11]. A. F. Fliri, W. T. Loging, P. F. Thadeio, and R. A. Volkmann, "Analysis of drug-induced effect patterns to link structure and side effects of medicines," Nature Chem. Biol., vol. 1, no. 7, 2005.

[12]. J. Scheiberet al., "Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis," J. Chem. Inf. Model., vol. 49, no. 2, 2009.

[13]. F. Wang, P.Zhang, N. Cao, J. Hu, and R. Sorrentino, "Exploring the associations between drug side-effects and therapeutic indications,"J.Biomed.Inform.,vol. 51, Oct.2014.

[14]. S. Mizutani, E. Pauwels, V. Stoven, S. Goto, and Y. Yamanishi, "Relating drug protein interaction network with drug sideeffects," Bioinformatics, vol. 28, no. 18, 2012.

[15]. Y. Yamanishi, E. Pauwels, and M. Kotera, "Drug side-effect prediction based on the integration of chemical and biological spaces," J. Chem. Inf.Model., vol. 52, no. 12, 2012.

[16]. M. Liu et al., "Determining molecular predictors of adverse drug reactions with causality analysis based on structure learning," J. Amer. Med. Inform.Assoc., vol. 21, no. 2, 2014.

[17]. F. Cheng et al., "Adverse drug events: Database construction and in silicoprediction," J. Chem. Inf. Model., vol. 53, no. 4, pp. 744-752, 2013.

[18]. W. Zhang, H. Zou, L. Luo, Q. Liu, W. Wu, and W. Xiao, "Predictingpotential side effects of drugs by recommender methods and ensemble learning," Neurocomputing, vol. 173, pp. 979-987, Jan. 2016.

[19]. Y.-G. Chen, Y.-Y.Wang, and X.-M.Zhao, "A survey on computational approaches to predicting adverse drug reactions," Current Topics Med.Chem., vol. 16, no. 30, 2016.

[20]. D. P. Williams and B. K. Park, "Idiosyncratic toxicity: The role of toxicophores and bioactivation," Drug Discovery Today, vol. 8, no. 22, pp. 1044-1050, 2003.