Classification and Optimization of chemical Drugs

Pradip A. Sarkate¹, Prof. A. V. Deorankar²

¹P.G. Student, Department of Computer Sci. & Engg. Govt. College of Engineering, Amravati, India
²Associate Professor, Department of Computer Sci. & Engg., Govt. College of Engineering, Amravati, India

Abstract—In data mining, classification is supervised learning method that can be used to design model which are describing important data classes. Classification in chemical drugs or medicine field is distinct from other field. There are large numbers of chemical and biomedical drug dataset which are necessary to classify the drugs. K-Nearest Neighbor is simple efficient method used to classify chemical drug dataset. Optimization technique is used to optimal solution, which are enhanced knowledge based classification. It will assign weigh value based on gradient descent algorithm which has better influenced of classification model.

Index Terms—Chemical medicine classification, data mining, k-NN algorithm, Optimization, Gradient descent algorithm.

I. INTRODUCTION

The drug database or drug information is publicly available on internet; Pharmacodia is a big platform that focuses on pharmaceutical research and development of drug. . The pharmaceutical industries control drug price during patent protection period which cost is very high to developing the new drug. While in market enter a new drug, they are approved by the State Food and Drug Administration (FDA). Drug information includes drug name, indication, physical and chemical properties of drugs.

The classification procedure consists of description classification algorithms which calculated the similarity based on sample datasets. K nearest neighbor is one of the simplest classification algorithms, K-NN classification algorithm used to calculate similarity between drugs sample data set which are classify different categories of class. In drug dataset optimization technique is used to find the optimal solution, the chemical drug has different properties and characteristics of drug which used to find the drug optimal solution. Gradient descent algorithm [12] is an optimization algorithm used to find value of weighed drugs parameter as function that can be minimized cost function.

II. PROPOSED WORK

The drug data used in experiment includes drug name, indication, physical and chemical properties of drugs including molecular mass, hydrogen bond donor, hydrogen bond acceptor, flexible rotation keys, polar surface area and hydrophobic constant.

A. Classification of chemical drugs:

Classification of chemical drugs or medicines we proposed k NN classification algorithm to categories the drug data set. We classify the drug or medicines with distance using k-NN algorithms. We classify unknown type of drug and calculating the similarity between medicines.

Steps for k-NN algorithm:

Step1: Determine parameter K, the number of nearest neighbors.
Step 2: For each case in the target data set that is the set to be predicted, locate the K closest members (the K nearest neighbors) of the training data set. An Euclidean Distance measure is used to calculate how close each member of the training set is to the target row that is being examined.
Step 3: Sort the distance and determine nearest neighbors based on the K-th minimum Distance.
Step 4: Gather the categories of the nearest neighbors.
Step 5: Use simple majority of the category of nearest neighbors as the prediction value of the new query instance.
Step 6: Repeat this procedure for the remaining cases in the target set.

The drug data which represent mathematical description of drugs data defined as X denote the vector of the chemical properties of the drugs, where each drug represents a sample. Thus, X={x1, x2, x3, x4, x5, x6}, where x1 denote molecular mass, x2 denotes hydrogen bond donors, x3 denotes bond
acceptors, x4 denotes flexible rotation key, x5 denotes polar surface area, and x6 denotes hydrophobic constant. There are n samples in the drug dataset X={\( (x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n) \)}, where each sample is a vector \((x_n,y_n)\). \(x_n\) denote the chemical properties of drugs and \(y_n\) denote the category of drugs. Suppose \(y_n^l\) is the category in \(y_n\) which has value of drug category, and \(y_n^p\) is the category in \(y_n\) which no value of category. Then the drug datasets can be divided into two independent subsets \(x_l\) and \(x_p\), where \(x_l=\{(x_1^l, y_1^l), (x_2^l, y_2^l), \ldots, (x_n^l, y_n^l)\}\) and \(x_p=\{(x_1^p, y_1^p), (x_2^p, y_2^p), \ldots, (x_n^p, y_n^p)\}\). It will train the classification model from the dataset \(x_l\) to predict the category of drug in \(x_p\), and then provide predictive value for the null values in \(y_n^p\).

B. Optimization of chemical drugs:
Classification of chemical drug based on physical and chemical properties of drugs which is used to calculating the distance between the drugs. The optimization technique is improved the classification method which are minimized the distance. The k-NN algorithm \(X\) denotes the vector of drugs which comparing other drugs of \(Y\) which are calculating the distance such as:

\[
\text{Distance} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_6 - y_6)^2}
\]

Optimization of chemical drug we used weigh parameter which has better influenced of classification system. Optimization technique which is involved partitioning the feature space and only computing distances within specific nearby volumes. The optimization distance is:

\[
\text{Optimized Distance} = \min\{w_1(x_1 - y_1)^2 + w_2(x_2 - y_2)^2 + \ldots + w_6(x_6 - y_6)^2\}
\]

In optimization technique is used to gradient descent method [12] to find the optimal solution. O'Donoghue which is used the following formula to calculate the drugs weight value [13].

\[
\phi = \frac{2}{1 + \sqrt{1 + \frac{4}{\phi^2}}}
\]

Based on the distance between the drugs, the k-NN algorithm is then used to classify them. We calculate the distance between all existing classified drugs from the training dataset.

III. RESULT
We are applying the classification model to the test data based on k-NN algorithm, we obtained the final algorithm for classification of drugs. The fig. shows the accuracy of classification of chemical drugs.

![Fig. Classification of drugs: (a) drugs for anti-tumors, (b) drugs for cancer (c) drugs for all other.](image)

From the drug dataset, we identified drugs with indications that related to Anti-Infection, Anti-Tumor, Cancer and Ulcer, and all other drugs such as Eye Disorder, Dyspepsia, and Schizophrenia etc. The drugs are classify three categories first category are Anti-Infection, Anti-Tumor, second category are Cancer and Ulcer and the third category are all other drugs such as Eye Disorder, Dyspepsia, and Schizophrenia etc. In fig. shows classification accuracy of drugs are specified the first category of drugs are 82%, second category of drugs are 81%, and third category of drugs are 84% respectively.

IV. CONCLUSION
There are large number of drugs which are necessary to classify different categories of drugs, thus we proposed the new k-NN classifier model. Optimization technique is used to find optimal solution, which are enhanced knowledge based classification. The classification system which is classified various drugs with known physical and chemical properties according to k-NN classification model. Based on properties of drugs, the optimization technique are calculated drug weight values to find optimal solution. The optimization of drugs which is better influenced of a classification model, thus classification of k-NN algorithm has good balance between accuracy.

REFERENCES


