Expert Systems with Applications 42 (2015) 4265–4273

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction

Seokho Kang [a], Pilsung Kang [b], Taehoon Ko [a], Sungzoon Cho [a,*], Su-jin Rhee [c], Kyung-Sang Yu [c]

[a] Department of Industrial Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Republic of Korea
[b] School of Industrial Management Engineering, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 136-701, Republic of Korea
[c] Department of Clinical Pharmacology and Therapeutics, Seoul National University College of Medicine and Hospital, 101 Daehak-ro, Jongno-gu, Seoul 110-744, Republic of Korea

ARTICLE INFO

Article history:
Available online 2 February 2015

Keywords:
Support vector machines
Ensemble
Data selection
Type 2 diabetes
Drug failure prediction

ABSTRACT

The treatment of patients with type 2 diabetes is mostly based on drug therapies, aiming at managing glucose levels appropriately. As the number of patients with type 2 diabetes continually increases worldwide, predicting drug treatment failure becomes an important issue. Support vector machine (SVM) can be a good method for the anti-diabetic drug failure prediction problem; however, it is difficult to train SVM on large-scale medical datasets directly because of its high training time complexity $\mathcal{O}(N^3)$. To address the limitation, we propose an efficient and effective ensemble of SVMs, called E^3-SVM. The proposed method excludes superfluous data points when constructing an SVM ensemble, thereby yielding a better classification performance. The proposed method consists of two phases. The first phase is to select the data points that are likely to be the support vectors by applying data selection methods. The second phase is to construct an SVM ensemble using the selected data points. We demonstrated the efficiency and effectiveness of the proposed method using the real-world dataset of the anti-diabetic drug failure prediction problem for type 2 diabetes. Experimental results show that the proposed method requires less training time to achieve comparable success, compared to the conventional SVM ensembles. Moreover, the proposed method obtains more reliable prediction results for each independent run of constructing an ensemble. In conclusion, firstly, the proposed method provides an efficient and effective way to use SVM for large-scale datasets. Secondly, we confirmed the suitability of SVM for the anti-diabetic drug failure prediction problem with an accuracy of about 80%.

1. Introduction

Diabetes is one of the most prevalent chronic diseases today. As the number of people with diabetes continually grows worldwide, the importance of research on the treatment of diabetes is progressively increasing. Particularly, type 2 diabetes is the most common, accounting for 85–90% of diabetes (Bennett, Guo, & Dharmage, 2007). Most patients with type 2 diabetes are under medical care with mono- or combination therapy of oral hypoglycemic agents, aiming at lowering glucose level. Glycated hemoglobin (HbA_{1c}) is an effective and widely used measurement of glucose level for patients with type 2 diabetes (Bennett et al., 2007; Lu, Walker, O'Dea, Sikaris, & Shaw, 2010). According to the guideline of the American Diabetes Association (ADA) (American Diabetes Association, 2014), combination therapy is recommended for the patients with type 2 diabetes who cannot be controlled by monotherapy. In addition, the ADA recommends an HbA_{1c} level of 7% or lower as the reasonable glycemic goal for most individuals.

Unfortunately, the majority of patients fail to achieve their glycemic goals (Brown, Nichols, & Perry, 2004). This is because the outcome of the diabetic treatment is highly related to various factors. The efficacy of anti-diabetic drugs can be affected by the characteristics of patients such as age, gender, obesity, and blood pressure. Moreover, the efficacy can also be affected by the interaction of various drugs. Therapies for type 2 diabetes are generally based on the combination of 2–3 oral hypoglycemic agents in order to obtain better and more reliable results (American Diabetes Association, 2014; Yki-Järvinen, 2001) In addition, because diabetes often leads to complications, drugs to treat complications are also administered to diabetic patients.

Predicting drug treatment failure is an important issue in the medical domain. Many studies have been conducted based on statistical analyses. However, it is difficult to predict the failure accurately using only statistical analyses because the failure is related to a variety of factors. Presently, the effectiveness of

* Corresponding author. Tel.: +82 2 880 6275.
E-mail addresses: prokids@snu.ac.kr (S. Kang), pilsung.kang@gmail.com (P. Kang), hooni915@snu.ac.kr (T. Ko), zoon@snu.ac.kr (S. Cho), rheesjin@snu.ac.kr (S.-j. Rhee), ksyu@snu.ac.kr (K.-S. Yu).

http://dx.doi.org/10.1016/j.eswa.2015.01.042
0957-4174/

machine leaning approaches for disease diagnosis has been reported by several studies (Hu, Wu, Lo, & Tai, 2012; Sajda, 2006; Zeng & Liu, 2010), and more recently, some researchers have attempted to apply machine learning approaches to diabetes (Huang, McCullagh, Black, & Harper, 2007; Marinov, Mosa, Yoo, & Boren, 2011). Most of them focus on prediction at the disease level. Machine learning approaches can also be effective for predicting drug treatment failure. However, to the best of our knowledge, there are relatively few efforts at predicting drug treatment failure using these approaches. The problem of anti-diabetic drug failure prediction can be defined as a classification problem, and has the characteristics of multivariate and complex relationships. Therefore, support vector machine (SVM) can be a good candidate as a classification algorithm.

SVM (Vapnik, 1995) is one of the most popular state-of-the-art classification algorithms, and shows superior generalization performance based on structural risk minimization principle. The effectiveness of SVM has been verified in various applications such as text categorization, handwritten digit recognition, image segmentation, and financial forecasting (Burges, 1998). Moreover, SVM is also known to be very effective in the medical domain (Barakat, Bradley, & Barakat, 2010; Yu, Liu, Valdez, Gwinn, & Khoury, 2010).

However, training of an SVM becomes a difficult problem when the size of a given dataset N is very large because the SVM takes $\mathcal{O}(N^3)$ of its training time complexity Kang and Cho (2014). the training of SVM involves solving a quadratic programming (QP) problem of $\mathcal{O}(N^3)$ complexity. Therefore, it is practically undesirable to train the SVM for a large-scale dataset directly. Commonly used approaches to alleviate the complexity are improving the efficiency of the QP process (Fan, Chen, & Lin, 2005; Joachims, 1999; Platt, 1999) and reducing the number of training data points by eliminating non-support vectors before the QP process (Li & Maguire, 2011; Shin & Cho, 2007).

When these approaches are insufficient, the training time of SVM can be further reduced by constructing an ensemble of SVMs that are trained with small bootstrap samples (Kim, Pang, Je, Kim, & Bang, 2003). The two typical ensemble methods, Bagging (Breiman, 1996) and Boosting (Freund & Schapire, 1997), can be employed to construct SVM ensembles (Kim et al., 2003; Wang et al., 2009). By doing so, we can obtain comparable classification accuracy by aggregating SVMs properly, although the classification accuracy of each SVM is lowered. One major concern is that a bootstrap sample might contain lots of superfluous data points when the size of such sample is set to very small, thereby resulting in the training of an ill-formed SVM.

In this paper, we propose an efficient and effective ensemble of SVMs for large-scale datasets based on data selection methods, called E^3-SVM. The proposed method is based on the fact that the SVM only uses support vectors to determine the decision boundary. In the proposed method, a reduced dataset is constructed by applying data selection methods (Shin & Cho, 2007; Li & Maguire, 2011) that select data points that are more likely to be the support vectors. That is, non-crucial data points of the original dataset are excluded in the reduced dataset. The ensemble of SVMs is constructed using bootstrap samples drawn from the reduced dataset. Consequently, the classification accuracy of the ensemble is improved by reducing the risk of using superfluous data points when training SVMs with small bootstrap samples. We investigated the efficiency and effectiveness of the proposed method through experiments on the anti-diabetic drug failure prediction problem.

The rest of this paper is organized as follows. In section 2, we briefly review the related work. In section 3, we describe our proposed method and section 4 reports the experimental results on the anti-diabetic drug failure prediction problem. The conclusion and future work are given in section 5.

2. Backgrounds

2.1. Support vector machines

SVM (Vapnik, 1995) seeks to find the maximum margin hyperplane $\mathbf{w}^T\varphi(\boldsymbol{x}_i)+b$ that separates the positive datapoints from negative datapoints. Given a training dataset $\mathcal{D}=\{\boldsymbol{x}_i,y_i\}_{i=1}^N$, where N is the number of training datapoints, $\boldsymbol{x}_i$ is an input feature vector and $y_i\in\{-1,1\}$ is the corresponding target class label, an SVM can be formulated as the following optimization problem:

$$\begin{aligned}\underset{\mathbf{w},b,\xi_i}{\text{minimize}}\quad & \frac{1}{2}\mathbf{w}^T\mathbf{w}+C\sum_i\xi_i\\ \text{subject to}\quad & y_i(\mathbf{w}^T\varphi(\boldsymbol{x}_i)+b)\geqslant 1-\xi_i,\\ & \xi_i\geqslant 0, i=1,\ldots,N,\end{aligned}\tag{1}$$

where $C>0$ is the parameter that controls the tradeoff between the training errors and the model complexity, ξ_i are slack variables used to achieve a soft margin, and φ is a non-linear mapping from an input space into a feature space. By introducing the Lagrange multiplier α_i, a corresponding dual problem can be derived as the following quadratic programming (QP) problem:

$$\begin{aligned}\underset{\alpha_i}{\text{maximize}}\quad & -\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)+\sum_i\alpha_i\\ \text{subject to}\quad & \sum_i\alpha_i y_i=0,\\ & 0\leqslant\alpha_i\leqslant C, i=1,\ldots,N,\end{aligned}\tag{2}$$

where k is a kernel function $k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\langle\varphi(\boldsymbol{x}_i),\varphi(\boldsymbol{x}_j)\rangle$, e.g., radial basis function (RBF) kernel $k(\boldsymbol{x}_i,\boldsymbol{x}_j)=exp(\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/2\sigma^2)$. Once the dual QP problem is solved, the resulting decision function at any test datapoint x is as follows:

$$f(\boldsymbol{x})=\mathbf{w}^T\varphi(\boldsymbol{x})+b=\sum_{i=1}^N\alpha_i y_i k(\boldsymbol{x}_i,\boldsymbol{x})+b=\sum_{i\in SV}\alpha_i y_i k(\boldsymbol{x}_i,\boldsymbol{x})+b.\tag{3}$$

Only those data points for which α_i is nonzero are referred to as support vectors, and they define the decision function. In the test phase, we estimate the class of the test datapoint $\boldsymbol{x}$ based on $sign(f(\boldsymbol{x}))$.

2.2. SVM Ensembles

An ensemble of classifiers builds a set of base classifiers and aggregate their outputs to classify new data points (Rokach, 2010). It often gives better accuracy and stability than any single classifiers. Kim et al. (2003) and Wang et al. (2009) employed SVMs to construct ensembles and confirmed the effectiveness. The two most famous ensemble methods, Bagging (Breiman, 1996) and Boosting (Freund & Schapire, 1997), can be used for SVM ensembles. Note that the best ensemble method can differ depending on the characteristics of problems or datasets.

2.2.1. Bagging

Bagging (Breiman, 1996), short for bootstrap aggregating, is a simple but powerful ensemble method to improve the accuracy and stability of learning algorithms. In Bagging, each bootstrap sample is drawn independently by random sampling with replacement. Several SVMs are trained independently on the bootstrap samples. The trained SVMs constitute an ensemble. The pseudo-code of Bagging for SVMs is presented in Algorithm 1.

Once the ensemble is constructed, final classification for new data points is done by certain aggregation rules such as majority voting, least squares estimation-based weighting, and double layer hierarchical combining (Kim et al., 2003). The most typical rule for