|  |  |  |  |
| --- | --- | --- | --- |
| Paper | Description | Pros | Cons |
| Machine learned job recommendation AUTHORS: Ioannis Paparrizos, B. Barla Cambazoglu, Aristides Gionis | The highly competitive and dynamic nature of the job market as well as personal preferences and goals lead individuals to change their jobs frequently in their lives. Moving to a new job, however, is not an easy decision, which may depend on many factors, such as salary, job description, and geographical location. These patterns may involve features extracted from the business profiles of employees, the profiles of institutions, and the job transitions themselves. In this paper, the authors address the problem of recommending suitable jobs to people who are seeking a new job and have formulated it as a supervised machine learning problem. Their technique exploits all past job transitions as well as the data associated with employees and institutions to predict an employee’s next job transition. They trained a machine learning model using a large number of job transitions extracted from the publicly available employee profiles in the Web. The results of their experiments demonstrate that the transition of an employee to an institution can be quite accurately predicted, significantly improving over a baseline predictor that always predicts the most frequent institution in the data. The results indicate that the most important feature in predicting a job transition is the current institution of the employee. | Uses a Decision table/Naive Bayes hybrid classifier. | Does not take into account the social aspects that could affect job transitions. |
| Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding AUTHORS: Juanjuan Wang; Mantao Xu; Hui Wang; Jiwu Zhang | Imbalanced data classification often arises in many practical applications in the context of medical pattern recognition and data mining. Most of the existing classification approaches are well developed by assuming the underlying training set is evenly distributed. However, they are faced with a severe bias problem when the training set is a highly imbalanced distribution thus leading to poor performance. SMOTE is an important approach by oversampling the positive class or the minority class. However, it is limited to an assumption that the local space between any two positive instances is positive or belongs to the minority class, which may not always be true in the case when the training data is not linearly separable. However, mapping the training data into a more linearly separable space, where the SMOTE algorithm can be conducted, can fix this problem. In this paper, the authors have combined Locally Linear Embedding algorithm (LLE) and SMOTE so that oversampling can be done on datasets that are non-linearly separable. Experimental results have demonstrated that this approach has better performance than traditional SMOTE. | Use of LLE algorithm along with SMOTE so that oversampling works on datasets that are non-linearly separable.  This paper experiments on 3 different datasets using 3 different classifiers. |  |
| Classification of Imbalanced Data by Combining the Complementary Neural Network and SMOTE Algorithm AUTHORS: Piyasak Jeatrakul, Kok Wai Wong, and Chun Che Fung | In recent years, many research groups have found that an imbalanced data set could be one of the obstacles for many Machine Learning algorithms. In the learning process of the ML algorithms, if the ratio of minority classes and majority classes is significantly different, ML tends to be dominated by the majority classes and the features of the minority classes are recognize slightly. As a result, the classification accuracy of the minority classes may be low when compared to the classification accuracy of the majority classes. The features in the minority classes are normally difficult to be fully recognized. In this paper, in order to re-balance the class distribution, the combined approaches of two techniques, Complementary Neural Network (CMTNN) and SMOTE, are proposed. While CMTNN is applied as an under-sampling technique, SMOTE is used as an over-sampling technique. CMTNN is used because of its special feature of predicting not only the "truth" classified data but also the "false" data. SMOTE is applied because it can create new instances rather than replicate the existing instances. | The proposed method is implemented on 4 different datasets, each with different number of attributes, to check for consistency.  Two different CMTNN variations are used. | In some cases, the CMTNN removes too many instances and the remaining instances are not enough for learning algorithms to obtain correct results. |
| Combination approach of SMOTE and biased-SVM for Imbalanced datasets AUTHORS: He-Yong Wang | Imbalanced data learning is problematic as traditional machine learning approaches fail to provide satisfactory results due to skewed class distribution. There are two solutions to this problem: increasing the number of minority class examples, called over-sampling, or decreasing the number of majority class examples, called under-sampling. A new approach to construct the classifiers from imbalanced datasets is proposed in this paper by combining SMOTE and BiasedSVM approaches. Often real-world data sets are predominately composed of normal examples with only a small percentage of abnormal examples. The cost of misclassifying an abnormal example into a normal example is often much higher than that of the reverse error. Experimental results confirms that the proposed combination approach of SMOTE and biased-SVM can achieve better classifier performance. | The proposed method is implemented on 4 different datasets, each with different number of attributes, to check for consistency. |  |
| A hybrid classifier combining SMOTE with PSO to estimate 5- year survivability of breast cancer patients AUTHORS : Kung-Jeng Wanga, Bunjira Makonda, Kun-Huang Chena, Kung-Min Wang | Data mining is a process to discover useful information through a large amount of data. This process is widely applied in medical, social science, management, engineering, and many other fields. In recent years, data mining is used for health care management to classify/justify disease prevalence and medical diagnosis. However, data mining problems are challenging in health care due to large, complex, heterogeneous, hierarchical time series data. The annual number of deaths caused by cancers is around million worldwide and breast cancer is one of the five most life threatening types of cancer. It is essential to know the survivability of the patients and to ease the decision making process regarding medical treatment and financial preparation. Meanwhile, false classification will cause wasted money and/or inappropriate treatments to cure the breast cancer. In this study, the authors propose a set of new algorithms to enhance the effectiveness of classification for 5-year survivability of breast cancer patients from a massive data set with imbalanced property. Results from this study show that the hybrid algorithm of SMOTE + PSO + C5 is the best one for 5-year survivability of breast cancer patient classification among all algorithm combinations. They conclude that, implementing SMOTE in appropriate searching algorithms such as PSO and classifiers such as C5 can significantly improve the effectiveness of classification for massive imbalanced data sets. | This paper tries implementing and comparing different combinations of algorithms and gives the best combination to be SMOTE + PSO+ C5 decision tree. |  |