پایان‌نامه‌ی کارشناسی ارشد: طاهره اندیشمند، ۱۳۹۶

کلاس‌بندی داده‌های ناقص برمبنای توابع باور و k نزدیک‌ترین همسایه

یکی از مسائلی که در زمینه طراحی داده¬ها می‌تواند بر پیچیدگی آنالیز بیفزاید، وجود داده¬های از دست رفته است. تعیین این مقادیر معمولاً در پردازش داده¬ها و یادگیری ماشینی از اهمیت زیادی برخوردار است و به همین دلیل در فرایند¬های پردازشی، سعی بر آن است که از تکنیک¬های خودکار برای بازیابی مقادیر گم شده استفاده شود. در این پایان نامه، راهکاری ترکیبی معرفی شده که مرکب از طبقه¬بندی مبتنی بر K نزدیکترین همسایگی و توابع باور استوار بر تئوری¬های دمپستر-شفر است و اولین مسئله¬ای را که مورد بررسی قرار داده، حل مشکل عدم قطعیت می¬باشد. با استفاده از تخمین مبتنی بر طبقه¬بند K نزدیکترین همسایگی، خروجی¬ها برآورد می¬شوند و در ادامه، نتایج تخمینی توسط K نردیکترین همسایگی با استفاده از تابع باور و انتخاب تابع جرم بهبود داده می¬شود. عامل وزن، نتیجه طبقه¬بندی مربوط به کلاس¬هاست که می¬تواند توسط مجموع وزن K انتخاب شده باشد و بردارهای تخمین برای سهیم شدن در برآورده کردن مقادیر از دست رفته را لحاظ نماید. مضاف بر این موارد، طبقه¬بندی از نتیجه تکرار پیاپی برای چند نوع داده که ذاتاً مقادیر از دست رفته به همراه دارند، صورت پذیرفته است. برای محک الگوریتم در دو مرحله طبقه¬بندی داده¬های همراه با مقادیر از دست رفته و نیز داده¬های بدون مقادیر از دست رفتگی، از 6 داده استفاده شده است. در مرحله نخست داده¬های بیماری کلیوی، سرطان دهانه رحم و بیماری هپاتیت تحلیل شده¬اند که با درصدهای متفاوتی از دست رفتگی داده¬ها مواجه هستند و در نهایت دقت¬های مرحله آزمایش به ترتیب معادل با 96/55%، 92/23% و 83/71% حاصل آمدند. در مرحله دوم، با ساختن مقادیر از دست رفته در داده¬های بیماری Glass، تیروئید و Wine از 5% تا 50%، خروجی-های بازیابی شده تعیین شدند و خطای مطلق متوسط و خطای جذر میانگین مربعات محاسبه شدند. برای هر سه داده، این دو مقدار در مقایسه با تکنیک¬هایی چون شبیه¬ترین عنصر و K نزدیکترین همسایگی، خطاهای کمتری بودند که به تفکیک گزارش خواهند شد. بکارگیری این روش ترکیبی می¬تواند تا حد چشمگیری بر بهبود دقت طبقه¬بندی اثرگذار باشد و خروجی¬هایی را پیش¬بینی نماید که در مقایسه با خروجی¬های واقعی دارای خطای ناچیزی باشد.

کلیدواژه‌ها: داده‌های از دست رفته،Kنزدیکترین همسایگی، توابع باور، تئوری دمپستر-شفر و عدم قطعیت.

شماره‌ی پایان‌نامه: ۱۲۷۴۱۰۰۶۹۵۱۰۰۸
تاریخ دفاع: ۱۳۹۶/۰۶/۲۹
رشته‌ی تحصیلی: مهندسی کامپیوتر- نرم افزار
دانشکده: فنی و مهندسی
استاد راهنما: مهندس حسام حسن‌پور

استاد مشاور: مهندس یاسر علمی‌سولا

M.A. Thesis:

Classification of incomplete data based on belief functions and K-nearest neighbors

One of the issues in the design of data can add to the complexity of the analysis is the lack of data. Determining these values is of paramount importance in data processing and machine learning. Therefore, in processing processes, it is attempted to use automated techniques to recover missing values.
In this thesis, a hybrid approach is introduced Which is based on K-Nearest neighborhood And the belief functions are based on the Dempster- Shafer theory And the first problem that we are examining is solving the uncertainty problem.
Using estimates based on the K nearest neighbor, the outputs are estimated. In the following, the estimated results are improved by K's best neighboring neighborhood using the belief function and the mass function selection.
The weight factor is the result of the classification of the classes, which can be chosen by the total weight of K and Consider the vectors of the estimation to share the lost values. In addition to these cases, the classification of the result is repeated in series for several types of data that inherently have lost values.
For benchmarking of the algorithm, six data have been used in two stages of classification of data along with missing values and data with no loss values.
In the first stage, data on kidney disease, cervical cancer and hepatitis have been analyzed, with different percentages of data loss. Finally, the accuracy of the test stage was obtained at 96.55%, 92.23% and 83.71%, respectively.
In the second step, by making the lost values in the data of Glass, Thyroid and Wine disease from 5% to 50%, the recovered outputs were determined, and the absolute error and mean square error were calculated.
For all three data, these two values had fewer errors compared to techniques like the most similar element and K nearest neighboring which will be reported separately.
Applying this combination method can greatly improve the accuracy of classification and predict outcomes that have little or no error compared to actual outputs