



پایان‌نامه‌ی کارشناسی ارشد: طاهره اندیشمند، ۱۳۹۶

کلاس‌بندی داده‌های ناقص بر مبنای توابع باور و k نزدیک‌ترین همسایه

یکی از مسائلی که در زمینه طراحی داده‌ها می‌تواند بر پیچیدگی آنالیز بیفزاید، وجود داده‌های از دست رفته است. تعیین این مقادیر معمولاً در پردازش داده‌ها و یادگیری ماشینی از اهمیت زیادی برخوردار است و به همین دلیل در فرایندهای پردازشی، سعی بر آن است که از تکنیک‌های خودکار برای بازیابی مقادیر گم شده استفاده شود. در این پایان‌نامه، راهکاری ترکیبی معرفی شده که مرکب از طبقه‌بندی مبتنی بر K نزدیکترین همسایگی و توابع باور استوار بر تئوری‌های دمپستر-شفر است و اولین مسئله‌ای را که مورد بررسی قرار داده، حل مشکل عدم قطعیت می‌باشد. با استفاده از تخمین مبتنی بر طبقه‌بندی K نزدیکترین همسایگی، خروجی‌ها برآورد می‌شوند و در ادامه، نتایج تخمینی توسط K نزدیکترین همسایگی با استفاده از تابع باور و انتخاب تابع جرم بهبود داده می‌شود. عامل وزن، نتیجه طبقه‌بندی مربوط به کلاس‌هاست که می‌تواند توسط مجموع وزن K انتخاب شده باشد و بردارهای تخمین برای سهم شدن در برآورده کردن مقادیر از دست رفته را لحاظ نماید. مضاف بر این موارد، طبقه‌بندی از نتیجه تکرار پیایی برای چند نوع داده که ذاتاً مقادیر از دست رفته به همراه دارند، صورت پذیرفته است. برای محک‌الگوریتم در دو مرحله طبقه‌بندی داده‌های همراه با مقادیر از دست رفته و نیز داده‌های بدون مقادیر از دست رفتگی، از 6 داده استفاده شده است. در مرحله نخست داده‌های بیماری کلیوی، سرطان دهانه رحم و بیماری هیپاتیت تحلیل شده‌اند که با درصدهای متفاوتی از دست رفتگی داده‌ها مواجه هستند و در نهایت دقت‌های 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 داده‌های بیماری Glass، تیروئید و Wine از 5٪ تا 50٪، خروجی‌های 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 تکنیک‌هایی چون شبیه‌ترین عنصر و K نزدیکترین همسایگی، خطاهای کمتری بودند که به تفکیک گزارش خواهند شد. بکارگیری این روش ترکیبی می‌تواند تا حد چشمگیری بر بهبود دقت طبقه‌بندی اثرگذار باشد و خروجی‌هایی را پیش‌بینی نماید که در مقایسه با خروجی‌های واقعی دارای خطای ناچیزی باشد.

کلیدواژه‌ها: داده‌های از دست رفته، K نزدیکترین 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



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